DOCUMENT RESUME

ED 313 224 SE 051 032

AUTHOR Robitaille, David F.
TITLE Evaluation and Assessment in Mathematics Education. Science and Technology Education Document Series No. 32.
INSTITUTION United Nations Educational, Scientific, and Cultural Organization, Paris (France). Div. of Science, Technical and Environmental Education.
REPORT NO UNESCO-ED-89/WS/6
PUB DATE 89
NOTE 142p.; Selection of papers presented to Theme Group T4, "Evaluation and Assessment," at the International Congress on Mathematical Education (6th, Budapest, Hungary, July 27-August 3, 1988).
AVAILABLE FROM Publications Div., UNESCO, 7 place de Fontenoy, 75700 Paris, France (price not available).
PUB TYPE Speeches/Conference Papers (150) -- Reports - General (140)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Achievement Tests; *Cross Cultural Studies; *Educational Assessment; *Elementary School Mathematics; Elementary Secondary Education; *Mathematics Achievement; Mathematics Education; Mathematics Skills; *Mathematics Tests; National Programs; Program Evaluation; *Secondary School Mathematics
IDENTIFIERS *Second International Mathematics Study

ABSTRACT

This document contains 15 papers selected from 47 papers presented at a theme group of the Sixth International Congress on Mathematical Education. The papers included have been divided into four groups. The first group consists of a survey paper, which discusses the history of assessment and evaluation. The second section consists of six papers dealing with findings from the Second International Mathematics Study, which was conducted in some 20 countries in the early 1980s. The next group of papers focuses on national initiatives in evaluation in mathematics, and includes such topics as the Verification of Mathematical Ability in Compulsory Schooling project in Italy; the Assessment of Performance Unit (APU) in the United Kingdom; and the National Assessment of Education Progress (NAEP) in the United States. The final set of papers deal with a variety of topics, including evaluation of students' problem-solving activities, diagnostic assessment, and evaluation of students' understanding of selected concepts. (YP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED313224

BEST COPY AVAILABLE

Evaluation and Assessment in Mathematics Education

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

E. Sainz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

SE 051 032

Science and Technology Education Document Series

No. 1 Glossary of Terms used in Science and Technology Education. 1981 (English)

No. 2 Methodologies for Relevant Skill Development in Biology Education. 1982 (English)

No. 3 Nutrition Education: Curriculum Planning and Selected Case Studies. 1982 (English) (Reprint in Nutrition Education Series No. 4)

No. 4 Technology Education as part of General Education. 1983 (English and French)

No. 5 Nutrition Education: Relevance and Future. 1982 (English) (Reprint in Nutrition Education Series No. 5)

No. 6 Chemistry Teaching and the Environment. 1983 (English)

No. 7 Encouraging Girls into Science and Technology Education: Some European Initiatives. 1984 (English)

No. 8 Genetically-Based Biological Technologies. 1984 (English)

No. 9 Biological Systems, Energy Sources and Biology Teaching. 1984 (English)

No. 10 Ecology, Ecosystem Management and Biology Teaching 1984 (Reprint 1986) (English)

No. 11 Agriculture and Biology Teaching. 1984 (English)

No. 12 Health Education and Biology Teaching 1984 (English)

No. 13 The Training of Primary Science Educators - A Workshop Approach. 1985 (English)

No. 14 L'Economie sociale familiale dans le developpement rural. 1985 (French)

No. 15 Human Development and Evolution and Biology Teaching. 1985 (English)

No. 16 Assessment: A Practical Guide to Improving the Quality and Scope of Assessment Instruments. 1986 (English)

No. 17 Practical Activities for Out-of-School Science and Technology Education. 1986 (English)

No. 18 The Social Relevance of Science and Technology Education. 1986 (English)

No. 19 The Teaching of Science and Technology in an Interdisciplinary Context. 1986 (English)

No. 20 Mathematics for All. 1986 (English, French in press)

No. 21 Science and Mathematics in the General Secondary School in the Soviet Union 1986 (English)

No. 22 Leisure, Values & Biology Teaching. 1987 (English and French)

No. 23 Use of Sea and its Organisms 1987 (English)

No. 24 Innovations in Science and Mathematics Education in the Soviet Union. 1987 (English)

No. 25 Biology and Human Welfare. Case Studies in Teaching Applied Biology. 1988 (English)

No. 26 Sourcebook of Science Education Research in the Caribbean. 1988 (English)

No. 27 Pour un enseignement intégré de la science et de la technologie : trois modules. 1988 (French)

No. 28 Microbiological Techniques in School. 1988 (English)

No. 29 Games and Toys in the Teaching of Science and Technology. 1988 (English, French)

No. 30 Field Work in Ecology for Secondary Schools in Tropical Countries. 1988 (English)

No. 31 Educational Materials Linking Technology Teaching with Science Education: Technology in Life. 1988 (English)

Cover photos

1. Photo Unesco/Paul Almasy
2. Photo UNATIONS
3. Photo Unesco/D. Bahrman
4. Photo rights reserved

Science and Technology Education

Document Series No. 32

Evaluation and Assessment in Mathematics Education

Edited by
David F. Robitaille
University of British Columbia
Vancouver, Canada

A selection of papers presented to Theme Group - T4
"Evaluation and Assessment"
at the
Sixth International Congress on Mathematical Education
Budapest, Hungary
July 27 - August 3, 1988

**Division of Science
Technical and Environmental
Education**

UNESCO

ED-89/WS/6

Paris, 1989

Preface

Evaluation and asscssment in science and mathematics education is the topic of two Unesco titles: this resource document, and the forthcoming Volume III of 'Innovations in Science and Technology Education'.

In August 1988, The Sixth International Congress on Mathematical Education (ICME 6) brought together in Budapest, Hungary, some 2,500 mathematics educators from seventy-two countries. One of the Theme Groups was devoted to evaluation and assessment in mathematics education. From the forty-seven papers presented in the thirteen sessions of this theme group, a selection was made to constitute this resource document.

This document complements 'Mathematics for All', which also appeared in Unesco's Science and Technology Education Document Series (Volume 20). As Thomas Romberg reminds us in his survey paper, the goals of a 'mathematics for all' policy are different from what happens in mathematics classrooms, and evaluation and assessment quantify this difference. Part 2 of the document centres around the Second International Mathematics Study, which examined the mathematics curriculum from three points of view: the intended, the implemented, and the attained. National initiatives in evaluation and selected topics make up the final parts of the document.

Unesco wishes to express its appreciation to the editor, David Robitaille, to the twenty-two authors of papers, to the University of British Columbia for preparing the manuscript, and to ICME 6.

The views expressed in this report are those of the authors and not necessarily those of Unesco.

We welcome comments on the contents of this document, which should be sent to: Division of Science, Technical and Environmental Education, Unesco, Place de Fontenoy, 75700 Paris, France.

Introduction

In recent years, an increasing amount of international attention in the mathematics education community has been focussed on evaluation and assessment. The organizers of ICME-6 acknowledged the level of interest in this topic by including a Theme Group on evaluation and assessment in the conference program to provide a forum for international discussion of evaluation activities in mathematics education. The members of the panel responsible for arranging the work of the Theme Group included Antoine Bodin (France), Raimondo Bolletta (Italy), Desmond Broomes (Barbados), David Robitaille (Canada; Chief Organizer), Toshio Sawada (Japan), and Julia Szendrei (Hungary).

Over a four-day period, 13 sessions were scheduled for Theme Group T4, and 47 papers were accepted for presentation by scholars from 13 countries. Because of limitations of space, it has not been possible to publish all of the papers accepted for presentation; in fact, only 15 are included in this collection. Brief summaries of all the papers presented at ICME - 6 may be found in the official proceedings of the conference.

The papers included here have been divided into four groups. The first group consists of one paper, and that is the Survey Paper prepared for the conference by Tom Romberg. The second section consists of papers dealing with findings from the Second IEA Mathematics Study which was conducted in some 20 countries in the early 1980s. The next group of papers focusses on national initiatives in evaluation in mathematics, and includes several papers on this topic from the United Kingdom. The final set of papers deal with a variety of topics including evaluation of students' problem-solving activities, diagnostic assessment, and evaluation of students' understanding of selected concepts.

Preparing a set of papers for publication requires a considerable amount of work in the best of circumstances. When the papers are submitted by authors from 13 countries, in a variey of formats, and in varying degrees of readiness for publication, the task can assume very large proportions indeed. I have been fortunate to have highly skilled and very dedicated assistance in this task, and I am very grateful to all those who helped in any way. First of all, I would like to thank Lori Teichman, a student in the teacher-education program at the University of British Columbia, who transferred all of the papers into *Microsoft Word*™ for the *Macintosh* microcomputer. My thanks also go to Brian Kilpatrick, a technician in the Department of Mathematics and Science Education at U.B.C., who provided technical support and advice and Michael Howell-Jones, of Education Audio-Visual Services of the Faculty of Education at U.B.C., who produced the camera-ready version of the document using *Pagemaker* ™. A special word of thanks goes to my colleague, James Sherrill, for proof-reading the entire set of papers so conscientiously.

I would also like to thank Unesco, and particularly Ed Jacobsen, for agreeing to publish this set of papers. Finally, I would like to thank Nancy Sheehan, Dean of the Faculty of Education at the University of British Columbia, for providing the resources needed to have the papers prepared for publication.

David F. Robitaille

Vancouver
January, 1989

List of First Authors

Raimondo Bolletta
Centro Europeo dell' Educazione
Villa Falconieri
Frascati, Italia

Margaret Brown
Centre for Educational Studies
King's College London
University of London
London, United Kingdom

Michael Dirks
North Central High School and
Spokane Falls Community College
Spokane, Washington, U.S.A.

Douglas Edge
Althouse College of Education
University of Western Ontario
London, Ontario, Canada

Derek Foxman
Department of Mathematics
National Foundation for Educational Research
United Kingdom

Edward A. (Skip) Kifer
Department of Educational Theory and Practice
College of Education
University of Kentucky
Lexington, Kentucky, U.S.A.

Curtis C. McKnight
Department of Mathematics
University of Oklahoma
Norman, Oklahoma, U.S.A.

David Nevo
School of Education
Tel-Aviv University
Israel

David F. Robitaille
Department of Mathematics and Science Education
University of British Columbia
Vancouver, British Columbia, Canada

Thomas A. Romberg
Department of Curriculum and Instruction
University of Winsconsin
Madison, Wisconsin, U.S.A.

Jane O. Swafford
Department of Mathematics
Illinois State University
Illinois, U.S.A.

Kenneth J. Travers
College of Education
University of Illinois
Champaign, Illinois, U.S.A.

Dylan Williams
King's College London
University of London
London, United Kingdom

Richard G. Wolfe
Ontario Institute for Studies in Education
Toronto, Ontario, Canada

Table of Contents

1 **Part 1: Survey Paper**

3 Thomas A. Romberg: *Evaluation: A Coat of Many Colors*

19 **Part 2: Papers from the Second International Mathematics Study**

21 Kenneth J. Travers: *Curriculum-Linked Assessment: A Model Based on the Second International Mathematics Study*
25 Michael K. Dirks, David F. Robitaille, & John Leduc: *Calculus in the High Schools of the United States and Canada (Ontario)*
39 Edward A. (Skip) Kifer: *Organizational Factors of School Systems*
49 Curtis C. McKnight & Thomas J. Cooney: *Content Representation in Mathematics Instruction*
59 David F. Robitaille: *Teaching Practices Employed in the Teaching of Algebra and Geometry*
69 Richard G. Wolfe: *Identification and Description of Opportunity to Learn and Growth in Achievement*

79 **Part 3: National Initiatives in Evaluation**

81 Raimondo Bolletta: *The Curriculum of the Scuola Media since 1979*
89 Derek D. Foxman & Graham Ruddock: *The 1987 APU Surveys: Some Preliminary Results*
97 Jane O.Swafford, Edward A. Silver, & Catherine A.Brown: *Findings from the Fourth National Mathematics Assessment in the United States.*

105 **Part 4: Selected Topics in Evaluation**

107 Margaret Brown: *The Graded Assessment in Mathematics Project*
113 Douglas Edge: *An Informal Diagnostic Instrument for Algebra*
121 Derek D. Foxman and Lynn S. Joffe: *Assessing Problem Solving in Small Groups*
129 David Nevo: *Widening the Perspective of Program Evaluation*
135 Dylan Williams: *Assessment of Open-Ended Work in the Secondary School*

Part 1

Survey Paper

EVALUATION: A COAT OF MANY COLORS

Thomas A. Romberg

*"**EVALUATE**: to judge or determine the worth or quality of" Webster's New World Dictionary, 1985, p. 484*

Evaluation in education has evolved from an initial and single concentration on measurement of achievement in order to make judgments about students, to the current and growing interest in providing information to support policy and program decision making. To make those latter judgments, information from students about their mathematical achievement is usually used. Thus, in this paper both the methods of gathering information from students and the use of that information to make a variety of judgments are examined.

The assessment of student performance in schools has a long history. However, contemporary models for both the gathering of performance data and the use of the information for policy and program decision making have only evolved during the past quarter-century. The purposes of this survey paper are:

1. to relate the gathering of assessment data to educational decision making;
2. to trace the history of this evolution. The assessment history begins in the 19th century and the evaluation history in the 1930s. However, in both cases, the developments in the past decade are stressed;
3. to illustrate the strengths and weaknesses of two contemporary social policy evaluation models. These are: evaluations of the impact of new mathematics programs, and large-scale profile evaluations; and
4. to describe four recent trends in evaluation and assessment.

Although the history and trends in assessment and evaluation are not unique to school mathematics, the emphasis and examples in this paper are all on assessing mathematical performance and the use of that information in instructional and policy contexts. Also, the examples have been selected to reflect the variety of models, methods, and procedures used throughout the world.

The principal point which should be understood is that at present there is considerable disparity between theory and practice. Academic considerations about goals, decisions, methods of gathering information, and validity of that information are in sharp contrast to the political and practical expectations of many governments and administrators. What is possible differs from what is done.

Educational Decision Making

The following examples are given to illustrate the relationships between measures of achievement and the variety of situations in which that information is used to make judgment (hence, the title of this paper):

1. A student has decided to study biology and would like to know whether she has the prerequisite knowledge to enroll in a biometrics course.
2. The admissions committee of a tertiary institution must select 100 students from some 800 who have applied for an engineering program.
3. A teacher would like to grade students on how well they understood the chapter on simultaneous linear equations just completed.
4. An official in a Department of Education has been asked to provide a legislative committee with information about pupil performance in mathematics.
5. A publishing company is interested in developing text to teach specific concepts of statistics to students in middle school. It needs feedback from teachers about the adequacy of the materials (i.e., what things were successful and what things were not) so that improvements could be made.
6. A researcher interested in early cognitive development with respect to mathematics would like to assess preschool children's ability to handle certain mathematical relationships, such as comparison of two sets with respect to numerosity.
7. An employer is interested in the mathematical capability of job applicants.
8. An official must decide which students are to be admitted to academic high schools and which to technical schools.

The preparation of this paper was supported by the Ford Foundation, the U.S. Department of Education, and the University of Wisconsin-Madison. The opinions expressed in this paper are those of the author and do not necessarily reflect the views of the Ford Foundation, the U.S. Department of Education, or the Wisconsin Center for Education Research.

These examples are only a few of the typical judgment situations in which information from students about their mathematical performance are frequently faced. In addition, they reflect the diversity of judgment (qualification, selection, placement, diagnosis, grading, profiling, researching, and so forth) as well as the variety of personnel involved in those decisions (students, administrators, teachers, developers, employers, and researchers).

Based on these examples, I have assumed that information from students about their mathematical achievement is important; and that information should influence educational decisions. The scenarios cited here are but a few examples of the many decisions facing educators throughout the world. Whether achievement data as a source of information actually influences schooling decisions is a separate and distinct empirical question. Nevertheless, valid data about student achievement should be available and used when making many such decisions.

Also, we must ask: How should such information be elicited? The answer to this question is based on a second assumption. The methods of gathering information (how data is collected, from whom, and how it is aggregated, organized, and reported) depends on the decisions to be made.

From these assumptions and the examples given above, I believe three elements of the decision-making process should be considered.

1. The decisions must be specifically identified. Gathering information without an explicit purpose in mind wastes time and resources. Although it is now fashionable to create data bases under the assumption that having such data will be useful, it has been shown that such data bases are rarely used or of value unless the purposes for which the data are to be used have been considered first.

2. The implications of the judgments to be made (or the questions to be answered) must be examined. This involves considering error in measurement, the errors in judgment (both Type I and Type II) that one is willing to tolerate, and whether the decisions are irrevocable. Teachers may be willing to accept considerable measurement error when administering chapter tests because they can rely on other information to judge a student's progress; a developer may be willing to live with high judgment errors in the development of a new instructional unit; while an admissions committee should seek minimal measurement error in choosing which applicants to accept to a program.

3. The "unit" about which the decisions are to be made must be determined (individuals, groups, classes, schools, materials, research questions, etc.). It has long been common practice to test all students on every item in every test; data from individuals can then be aggregated at any group level for any purpose. This practice is extremely wasteful, both in terms of cost and time. For example, the administration of a standardized test merely to publish the results in the local press (as is common in the United States of America) is wasteful both of student time and the cost of administration and scoring. Profiling school performance can be accomplished more efficiently.

In summary, to assess student performance in mathematics, one should consider the kinds of judgments that need to be made and tailor the assessment procedures in light of considerations about those decisions. This is particularly important because the information is being used by policy makers to make programmatic decisions.

History of Assessment and Evaluation

The history of the measurement of human behavior, with primary reference to the capacities and educational attainments of school students, may be divided roughly into four periods. During the first period, from the beginning of historical records to about the 19th century, measurement in education was quite crude. During the second period, embracing approximately the 19th century, educational measurement began to assimilate from various sources the ideas and the scientific and statistical techniques which were later to result in the psychometric testing movement. The third period, dating from about 1900 to the 1960's, can be characterized as the psychometric period. The final period, dating from the 1960's to the present, is the policy-program evaluation period.

Early Examinations

The initiation ceremonies by which primitive tribes tested the knowledge of tribal customs, endurance, and bravery of young men prior to admission to the ranks of adult males may be among earliest examinations employed by human beings. Use of a crude oral test was reported in the Old Testament, and Socrates is known to have employed searching types of oral quizzing. Elaborate and exhaustive written examinations were used by the Chinese as early as 2200 B.C. in the selection of their public officials. These illustrations may be classified as historical antecedents of performance tests, oral examinations, and essay tests. However, there is no evidence that different individuals ever took the same tests, and all judgments were made by officials in a manner similar to examinations given to doctoral students today.

Educational Testing in the 19th Century

Three persons made outstanding contributions to 19th-century developments. The ideas of these men—Horace Mann, George Fisher, and J. M. Rice—appear to be forerunners of developments during the present century.

The first school examinations of note appear to be those instituted in the Boston schools of 1845 in the United States as substitutes for oral tests when enrollments became so large that the school committee could no longer examine all pupils orally. These written examinations, in arithmetic, astronomy, geography, grammar, history, and natural philosophy, impressed Horace Mann, then secretary of the Massachusetts Board of Education. As editor of the *Common School Journal*, he published extracts from them and concluded that the new written examination was superior to the old oral test in these respects:

1. It is impartial
2. It is just to the pupils.
3. It is more thorough than older forms of examination.
4. It prevents the "officious interference" of the teacher.
5. It "determines, beyond appeal or gainsaying, whether the pupils have been faithfully and competently taught."
6. It takes away "all possibility of favoritism."
7. It makes the information obtained available to all.
8. It enables all to appraise the ease or difficulty of the questions.

(Greene, Jorgenson, & Gerberich, 1953)

Although these ideas are those represented by modern tests, the instruments themselves were inadequate. However, in successive issues of the *Common School Journal*, Mann suggested most of the elements in examinations that are found in contemporary measurement (e.g., timed responses by students to identical questions).

To Reverend George Fisher, an English schoolmaster, goes the credit for devising and using what were probably the first objective measures of achievement. His "scale books," used in the Greenwich Hospital School as early as 1864, provided means for evaluating accomplishments in handwriting, spelling, mathematics, grammar and composition, and several other school subjects. Specimens of pupil work were compared with "standard specimens" to determine numerical ratings that, at least for spelling and a few other subjects, depended on errors in performance (Greene, Jorgenson, & Gerberich, 1953). Scoring procedures for many examinations still follow this procedure (e.g the English "O" -level exams).

The use of test information for program evaluation was first developed by J. M. Rice, an American dentist. In 1894, he developed a battery spelling test. Having administered a list of spelling words to pupils in many school systems and analyzed the results, Rice found that pupils who had studied spelling 30 minutes a day for eight years were not better spellers than children who had studied the subject 15 minutes a day for eight years. Rice was attacked and reviled for this "heresy," and some educators even attacked the use of a measure of how well pupils could spell for evaluating the efficiency of spelling instruction. They intended that spelling was taught to develop the pupils' minds and not to teach them to spell. It was more than ten years later that Rice's pioneering resulted in significant attention to objective models in educational testing (Ayres, 1918).

The Psychometric Period

This era began shortly after the turn of the century. Although the historical antecedents sketched in the preceding paragraphs were essential prerequisites, developments first in mental testing and shortly after in achievement testing are at the roots of this era.

General Intelligence Tests. Attempts to measure general intelligence, or ability to learn or ability to adapt oneself to new situations, had been made both in the United States of America and in France. The first individual test was developed in France, and the first group test was developed some years later in the United States of America.

Individual intelligence scales were originated in 1905 by Binet and Simon in France. Their first scale was devised primarily for the purpose of selecting mentally retarded pupils who required special instruction. This pioneer individual-intelligence scale was based on interpreting the relative intelligence of different children at any given chronological age by the number of questions of varied types and increasing levels of difficulty they could answer. These characteristics were all re-embodied in the 1908 and 1911 revisions of the Binet-Simon Scale and remain basic to most individual intelligence scales today. The 1908 revision introduced the fundamentally important concept of mental age (MA) and provided means for obtaining it (Freeman, 1930).

The first group intelligence test was Army Alpha, used for the measurement and placement of American army recruits and draftees during World War I. It was the product of the collaboration of various psychologists working on group intelligence tests when the United States entered the war.

Aptitude Tests. The measurement of aptitudes, or those potentialities for success in an area of performance that exist prior to direct acquaintance with that area, was closely related to intelligence testing. Early attempts to measure general intelligence tested many

specific traits and aptitudes, but that approach was abandoned after Binet showed that tests of more complex forms of behavior were superior. It was soon apparent, however, that general intelligence tests were not highly predictive of certain types of performance, especially in the trades and industries. Munsterberg's aptitude tests for telephone girls and streetcar motormen were followed by tests of mechanical aptitude, musical aptitude, art aptitude, clerical aptitude, and aptitude for various subjects of the high school and college curricula prior to 1930 (Watson, 1938). Spearman's (1904) splitting of total mental ability into a general factor and many specific factors had its influence on this movement.

Achievement Tests. Modern achievement testing was stimulated by Thorndike's (1904) book on mental, social, and educational measurements. Through his book and his influence on his students, Thorndike was predominantly responsible for the early development of standardized tests. Stone, a student of Thorndike's, published the first arithmetic reasoning test in 1908. Between 1909 and 1915, a series of arithmetic tests and scales for measuring abilities in English composition, spelling, drawing, and handwriting were published (Odell, 1930). Literally thousands of standardized achievement tests have been published during the last half-century.

The reasons for presenting this brief history of testing are threefold. First, what is referred to as the modern testing movement began with a selection problem (Binet & Simon) and a placement problem (Army Alpha). It was assumed that a single measure (e.g. MA) or index (e.g., IQ) could be developed to compare individuals on what was assumed to be a general, fixed, unidimensional trait. In turn, the procedures that evolved in developing and administering these tests were used in aptitude and achievement tests. Second, the testing procedures now considered typical in many countries were developed for group administration of early intelligence tests. Such tests are comprised of a set of questions (items), each having one unambiguous answer. In this sense, such tests are "objective" since subjective inferences are not necessary. All subjects are administered the same items under standard (nearly identical) situations with the same instructions, time, constraints, etc. Furthermore, subjects' answers can be easily scored as correct or not, the total number of correct answers tallied, tallies transformed, and transformed scores compared. Psychometrics involving the application of statistical procedures to such tests developed as a field of study in the 1920s.

Most importantly, it should be understood that the testing movement was a product of an historical era. It grew out of the machine-age thinking of the industrial revolution of the past century. Business, industry, and, in particular, schools have been conceived, modified, and operated based on this mechanical view of the world since before the turn of the century.

The Policy-Program Evaluation Period

Information about student achievement has long been used by teachers and educators to make decisions about students. However, the use of that information for wide-scale policy or program judgments is recent. It began with the burst of reform policies associated with the mid-60s Great Society initiatives in the United States. Federal-level insistence on evaluation of those initiatives was thrust upon a largely unprepared field. Little expertise existed in the agencies responsible for carrying out evaluations in areas as diverse as bilingual education, career education, compensatory programs, reading, or mathematics. In fact, in the United States the initial training institute on program evaluation was held at the University of Illinois in 1963 (Directed by Lee Cronbach).

That early work followed the notions of Ralph Tyler (1931), the "father of educational evaluation." His conception of evaluation involved comparison between intended and observed program objectives. Tyler's model of evaluation in education dominated until the 1970s when that approach, like traditional social science models, were found inadequate as guidance for policy and practice. That evaluation model was based on the hypothetico-deductive traditions of "hard science." It focused on outcomes, and sought "significant differences." Initial evaluations of federal education programs used experimental methodology to assess student achievement and program accomplishments. As applied, this approach paid little attention to the context of program activities or the processes by which program plans were translated into practice (Eash, 1985; O'Keefe, 1984). Talk about evaluation included fairly rigid rules for "good" design and "scientific" evaluation. In particular, they gathered data on student performance using standard achievement tests.

In summary, evaluation for policy and program purposes began in the 1960s by attempting to apply "scientific" principles using notions from experimental sciences. The information from students was from tests based on the psychometric assessment technique outlined above. Again, this approach to evaluation is a product of "industrial age" thinking.

Two Social Policy Evaluation Models

Policy makers (legislators, government officials, school administrators, ...) must make many decisions related to the teaching and learning of mathematics. In this section, two evaluation models often used by policy makers are examined in detail so that their strengths and weaknesses are apparent.

Program Evaluation

Attempts to evaluate the impact of new curriculum programs involved the comparison of the performance of a group of students who had studied mathematics from that curriculum with an alternate group (most often a non-equivalent group). Performance was measured from both groups based on scores derived from the same instrument. Initially, in the United States, standardized tests were used; later it became common to use criterion-referenced tests.

Norm-referenced standardized tests have become an annual ritual in most American schools. Such tests are designed to indicate a respondent's position in a population. Each test is comprised of a set of independent, multiple-choice questions. The items have necessarily been subjected to a preliminary trial with a representative pupil group so that it is possible to arrange them in the desired manner with respect to difficulty and the degree to which they discriminate among students. Also, the test is accompanied by a chart or table to be used to transform test results into meaningful characterizations of pupil mental ability or achievement (grade-equivalent scores, percentiles, stanines, etc.)

Three features of such tests merit comment. First, although each test is designed to order individuals on a single (unidimensional) trait, such as quantitative aptitude, the derived score is not a direct measure of that trait. Second, because individual scores are compared with those of a norm population, there will always be some high and some low scores. This is true even if the range of scores is small. Thus, high and low scores cannot fairly or accurately be judged as "good" or "bad" with respect to the underlying trait. Third, test items are assumed to be equivalent to one another. They are selected on the basis of general level of difficulty (p-value) and some index of discrimination (e.g., non- spurious biserial correlation). Furthermore, no claim is made that the items are representative of any well-defined domain.

The primary strengths of standardized tests are that they are relatively easy to develop, inexpensive, and convenient to administer. Furthermore, the results are comprehensible since standard procedures are followed. Their primary weakness is that they are often used for decisions they were not designed to address. For example, aggregating standardized scores for students in a class (school, district, etc.) to produce a class profile of achievement (class mean) is very inefficient. The tests provide too little information in light of the high cost involved. In fact it has become clear that such tests are of little value for most evaluations since the items are not selected as representative of the mathematical domains in the curriculum.

Unfortunately, in the United States their use appears to be more strongly related to political, rather than educational, uses. For example, it is claimed that elected officials and educational administrators increasingly use the scores from such tests in comparative ways—to indicate which schools, school districts, and even individual teachers give the appearance of achieving better results (National Coalition of Advocates for Students, 1985). Such comparisons are simply misleading. One can only conclude that standardized tests are unwisely overused.

Criterion-referenced tests are a product of the behavioral objectives movement in the 1960's. They were developed to provide teachers with an objective set of procedures with which to make instructional decisions. Item development was based on the identification of a set of such behavioral objectives as: "the subject, when exposed to the conditions described in the antecedent, displays the action specified in the verb in the situation specified by the consequent to some specified criterion" (Romberg, 1976, p. 23). Items randomly selected from a pool designed to represent the antecedent conditions and the same action verb are given to students. From their responses, diagnosis of problems or judgments of mastery of objectives can be made.

Three features of these tests should be mentioned. First, they usually are designed as part of a curriculum to be administered to individuals at the end of some instructional topic. Often, they are given individually, and teachers' judgments are made quickly. Second, they have occasionally been used in group settings. For example, the comprehensive achievement monitoring scheme (Gorth, Schriber, & O'Reilly, 1974) periodically assesses student performance on a set of objectives. Third, decisions about performance are made with respect to some *a priori* criteria.

The strengths of objective-referenced tests lie in their usefulness in instruction. As long as instruction on some topic focuses on the acquisition of some specific concept or skill, such tests can be used to indicate whether or not the concept has been learned or the skill mastered. Furthermore, such tests are scored easily and are readily interpretable.

Four weaknesses need to be discussed. First, the specification of a set of behavioral objectives fractionates mathematical knowledge. In no way is it possible to reflect the interrelatedness of concepts and procedures in any domain. Second, objective-referenced tests are costly to construct because hundreds of objectives are included in any instructional program. Third, simple aggregation across objectives is not reasonable since objectives are interdependent. Fourth, and most importantly, items for higher level or complex problem-solving processes are very difficult to construct and are usually omitted. In fact, as used, these tests reinforce the factory metaphor of schooling. They clearly do not

reflect how students reason about problem situations, interpret results, or build arguments.

The problem faced by most program evaluators in the 1960's was a residue of the "scientific" traditions. The only evidence deemed of value was student performance at the end of treatment when compared with that of an alternate treatment group, and the evidence was gathered from either a standardized test or later a criterion referenced test. The results of such examinations (e.g. The National Longitudinal Study of Mathematical Abilities, Begle & Wilson, 1970) did not show that the new program was uniformly superior to the old program, but rather that different curricula are associated with different patterns of achievement.

Policy Profiles

Profile tests are intended to provide information on a variety of mathematical topics so that policy makers can compare individuals or groups in terms of those topics. Profile tests have become very popular. They have been developed for several major studies of mathematical performance, including the National Assessment of Educational Progress (NAEP) in the United States, the First International Mathematics Study (FIMS), the Second International Mathematics Study (SIMS), and the Assessment of Performance Unit (APU) in England.

Five features of profile assessments distinguish them from prior tests. First, they make no assumption of an underlying single trait; rather, the tests are designed to reflect the multidimensional nature of mathematical content. Second, items similar to those used in standardized or criterion-referenced tests are used. However, it must be acknowledged that the mathematics profiles developed by the APU in England (Foxman et. al., 1980, 1981) differ from most other profile assessments in the choice and form of items or exercises administered. Their exercises include a variety of open-ended questions, performance tasks, etc. Third, the unit of investigation is a group rather than an individual. Matrix sampling is usually used so that a wider variety of items can be included. Fourth, comparisons between groups are shown graphically on actual scores so that no transformations are needed. (See, for example, Figures 1 and 2.)

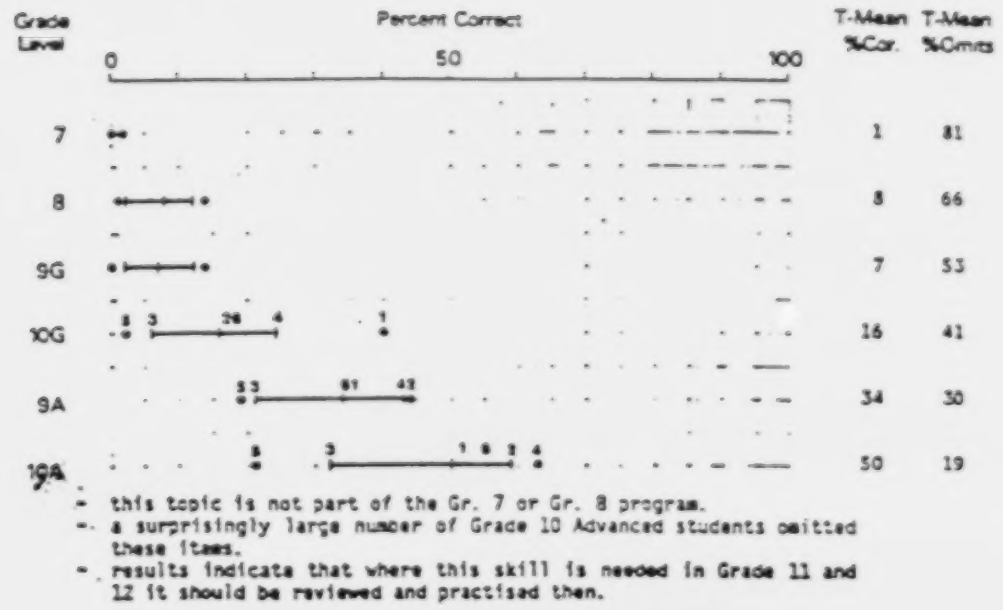
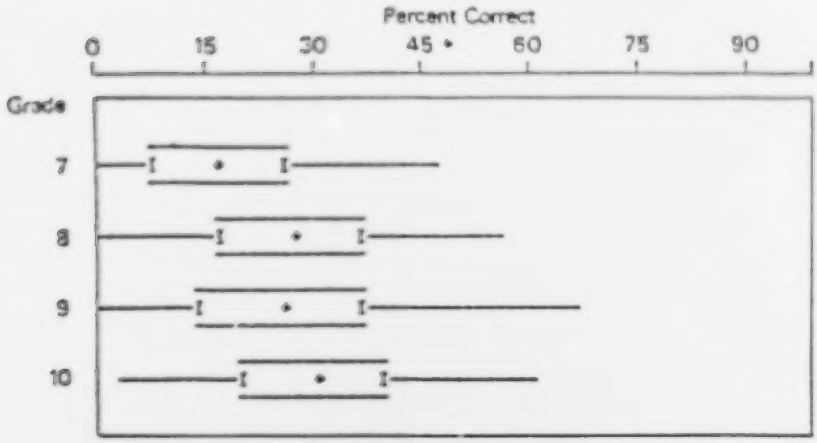


Figure 1. *Algebra - Equations and Inequalities*. Range of Correct Responses to the six Instruments, by Grade. (from McLean, 1982, p.207)



Statistical Summary

Grade Level	No. of Classes	Grade Mean	Grade St. Dev.
7	97	18.6	11.8
8	99	26.8	12.9
9	122	25.6	13.4
10	103	30.4	13.8

Figure 2. *Percentages.* Range of Correct Responses to Topic Group by Grade. (from Mclean, 1982, p.138)

Finally, validity is determined in terms of content and/or curricular validity. Mathematicians and teachers are asked to judge whether individual items reflect a content by behavior cell in a matrix. In fact, the usual approach in profile testing is to specify a content by behavior matrix. For example, to establish a framework for an item domain, a content by behavior grid was developed for each target population in SIMS (Weinzweig & Wilson, 1977). The content dimensions for both Grade 8 and Grade 12 populations were intended to cover all topics likely to be taught in any country. For Grade 8, the content outline contained 133 categories under five broad classifications: arithmetic, algebra, geometry, statistics, and measurement. For Grade 12, the content description was broader, containing 150 categories under seven headings: sets and relations, number systems, algebra, geometry, elementary functions and calculus, probability and statistics, and finite mathematics.

For each population in the SIMS study, the behavior dimension referred to four levels of cognitive complexity expected of students: computation, comprehension, application, and analysis. This classification is adapted from Bloom's taxonomy of educational objectives (1956). The adaptation involved replacing "knowledge" with "computation", and eliminating the higher levels of synthesis and evaluation. Data from such tests can then be reported in several ways. First, it can be reported in terms of items or cell means. For example, in Figure 1, the means are presented for six items on a topic (each given a different instrument) for different students at different grades in the province of Ontario, Canada (McLean, 1982). Second, item scores can be aggregated by columns to yield cognitive level scores or by rows to yield topic scores (see Figure 2).

A strength of profile achievement tests is that they can provide useful information about groups, and are particularly useful for general evaluations of changed educational policy that directly affects classroom instruction. However, profile achievement tests are weak in four specific areas. First, because they are designed to reflect group performance, they are not useful for individual ranking or diagnosis. An individual student takes only a sample of items. Second, they are somewhat more costly to develop, and harder to administer and score than prior tests. Third, because they yield a profile of scores, they are often difficult to interpret.

Finally, however, the primary weakness of most profile achievement tests center on the outdated assumptions underlying the two dimensions of content-

by-behavior matrices. The content dimension involves a classification of mathematical topics into "informational" categories. As I (1983) have argued:

> "Informational knowledge" is material that can be fallen back upon as given, settled, established, assured in a doubtful situation. Clearly, the concepts and processes from some branches of mathematics should be known by all students. The emphasis of instruction, however, should be "knowing how" rather than "knowing what" (p. 122).

Furthermore, items in any content category are tested as if they were independent of one another, a practice that ignores the interconnections between ideas within a well-defined mathematical domain. Schoenfeld and Herrmann (1982) cautioned about the problems inherent in testing students on isolated tasks.

> If they succeed on those problems, we and they congratulate each other on the fact that they have learned some powerful mathematical techniques. In fact, they may be able to use such techniques mechanically while lacking some rudimentary thinking skills. To allow them, and ourselves, to believe that they understand the mathematics is deceptive and fraudulent (p. 29).

Thus, the items should reflect the interdependence (rather than independence) of ideas in a content domain.

The behavior dimension of matrices has always posed problems. All agree that Bloom's Taxonomy (1956) has proven useful for low-level behaviors (knowledge, comprehension and application) but difficult for higher levels (analysis, synthesis, and evaluation). Single-answer, multiple-choice items are not reasonable at those levels. One problem is that the Taxonomy suggests that "lower" skills should be taught before "higher" skills. The fundamental problem is the Taxonomy's failure to reflect current psychological thinking, and the fact that it is based on "the naive psychological principle that individual simple behaviors become integrated to form a more complex behavior" (Collis, 1987, p. 3). In the past 30 years, our knowledge about learning and how information is proccessed has changed and expanded.

In summary, profiling is important but current profile tests fail to reflect the way mathematical knowledge is structured or how information is processed within mathematical domains.

Trends

In this section four trends are described. The first three are academic or theoretical trends apparent in the literature on assessment and evaluation. The last is a conservative political and practical trend which, in some respect, runs counter to the other trends.

The Trend in Program Evaluation

Far from the limited alternatives of "treatment/control" or randomized designs (see Campbell & Stanley, 1966), contemporary evaluators have developed a diverse assortment of evaluation approaches from which to choose, given purpose, context, program stage, etc. In contrast to the 1960's "one right way" today evaluators have multiple (and not always compatible) approaches. This trend began in the 1970's when scholars trained in disciplines other than experimental psychology were asked to assist in educational evaluations. Scholars like Michael Young (1975), Michael Apple (1979), and Tom Popkewitz (1984), whose training was in anthropology, sociology, and political science, brought the methods of information gathering and analysis of those disciplines to evaluation. In fact, the list of names of designations for the new methods and models can be confusing to someone unfamiliar with the field of evaluation and the controversies that underlie the various empirical procedures. For example, the catalogue of choices now available to evaluators includes: goal-free evaluation (Scriven, 1974); advocate evaluation (Stake & Gjerde, 1974; Reinhard, 1972); connoisseurship (Eisner, 1976); user-driven evaluation (Patton, 1980); ethnographic evaluation (Fetterman, 1984); responsive evaluation (Stake, 1974); naturalistic inquiry (Guba & Lincoln, 1981).

These diverse approaches to evaluation differ on many dimensions. Chief among them are the role of the evaluator (from educator to management consultant to assessor to advocate), role of the client (from active stakeholder and collaborator to passive recipient of evaluation product), to overall design (from experimental or quasi-experimental to exploratory), and focus (on process—formative evaluation—or outcome—summative evaluation). Each of these dimensions corresponds to the contingencies upon which evaluation choices are based: purpose, decision context, stage of program development, status of theory or knowledge base, etc. One consequence for product development was the specification of four stages of evaluation : 1) product design stage— this involves developing a needs assessment; 2) product creation stage — this involves gathering formative data to improve the product; 3) product implementation stage — this involves demonstrating differences between products and making sure appropriate support services are available; and 4) product illuminative stage—this involves an in-depth examination of how the product is actually used (Romberg, 1975).

Another consequence has been the use of a convergent strategy: i.e. using several different evaluation models with the same program. For example, in the

IGE Evaluation Study which I directed (Romberg, 1985), we gathered data about reading and mathematics in schools in four phases. Phase 1 involved large-scale survey procedures (including the use of a standardized test). Phase 2 was a follow-up study examining the validity of the Phase 1 data. Phase 3 was an ethnographic study of six exemplary IGE schools. Finally, Phase 4 was a detailed examination in Grades 2 and 5 using time-on-task observations and the repeated administration of criterion-referenced tests.

Note also that evaluation experts began calling for better and different instrumentation to gather information about student performance. Overall, while program evaluation models have proliferated and the questions which they must address have become clear, the information used to answer questions too often still comes from inappropriate tests.

It is only recently that it has become apparent that the kind of evidence one needed to gather to judge many programs is, of necessity, different from that obtained from conventional assessment procedures. Tests given in a restricted format (eg. multiple-choice items) and in a restricted time fail to capture the important aspects of doing mathematics. During the past decade researchers have developed a plethora of procedures for gathering information from students: think-aloud interview procedures, performance tasks, projects (both individual and group), hierarchical reasoning tasks, etc. Unfortunately, with one notable exception, these procedures, because of cost of administration, have not been used in program evaluations.

The exception is the evaluation of the Hewet Mathematics A Project in the Netherlands (deLange, 1987). In that evaluation five different tasks were used to gather information: timed written tests, two-stage tasks, take-home task, an essay task, and an oral task. Together the overall picture of how well students learned from that program is greatly enriched as a result of using information from the five tasks than would have been possible using any one.

Trends in External Assessment

While past assessment procedures are useful for some purposes and undoubtedly will continue to be used, they are products of an earlier era in educational thought. Like the Model T Ford assembly line, objective tests were considered an example of the application of modern scientific techniques in the 1920s. Today, we are both technologically and intellectually equipped to improve on outdated methods and instruments. The real problem is that all three forms of tests (profile, standardized, and criterion-referenced) are based on the same set of assumptions: an essentialist view of knowledge, a behavioral theory of learning, and a dispensary approach to teaching. It should be obvious that new assessment techniques need to be developed which are consistent with a different view of knowledge, learning, and teaching.

New evaluation models are being developed which demand new assessment procedures. One new approach is based on the specification of mathematical domains and the development of items that reflect that domain (Romberg, 1987). In turn, this assessment approach grows out of the extensive research on such domains. A good example is the work of Gerard Vergnaud with respect to "conceptual fields" (cf. 1982). The principles that are followed in this approach include:

> **Principle 1.** A set of specific and important mathematical domains need to be identified, and the structure and interconnectedness of the procedures, concepts and problem situations in each of the domains would need to be specified.

Note that this approach is different from the current approach to specifying the mathematical content of a test in that networks are being defined rather than categories. This means that the interconnections of concepts and procedures with problem situations are as important as mastery of any node (e.g. a specific procedures). For example, consider the following exercise in second grade addition and subtraction:

> Sue received a box of candy for her birthday. She shared 27 pieces with her friends and now has 37 pieces left. How many pieces of candy were originally in the box?

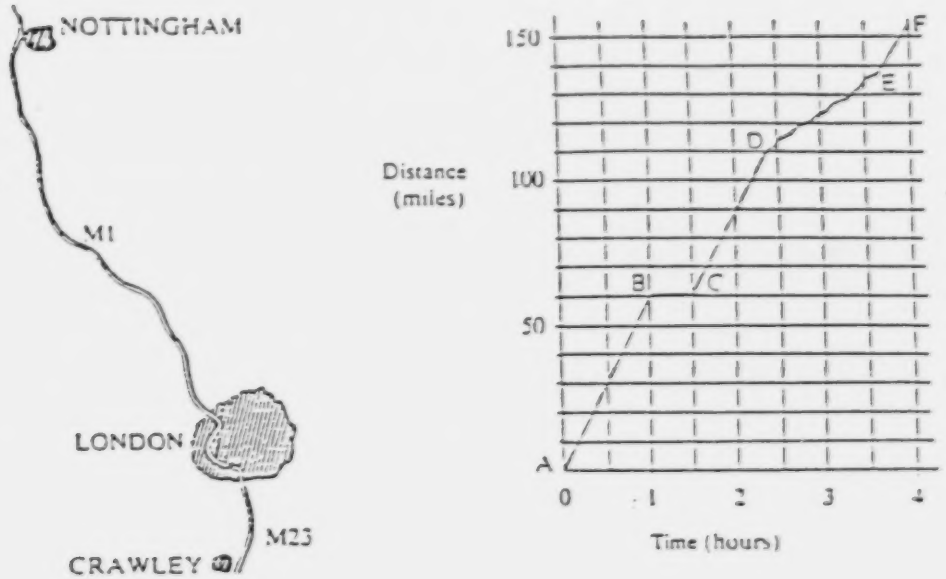
To solve this exercise, a child would be expected first to represent the quantitative information with the subtraction sentence $\square - 27 = 37$. Second, the sentence should be transformed to the addition sentence $27 + 37 = \square$; then the addition performed to yield an answer. What is important is that the child must know that separating situations can be represented by subtraction sentences, that subtraction sentences can be transformed into equivalent addition sentences, and that there are procedures for performing additions, etc. Each piece of knowledge, while important, contributes to a solution process or way of reasoning about a situation that is more important than any single concept or process.

> **Principle 2.** A variety of tasks should be constructed that reflect the typical procedures, concepts, and problem situations of that domain.

This is the key principle in that the envisioned tasks are not just a more clever set of paper-and-pencil, multiple-choice test items. Although some typical test items may be appropriate for determining mastery of

some specific concept or process, many of the tasks must be different. For example, some should be exercises which require the student to relate several concepts and procedures such as the example from addition and subtraction given above. Some would ask students to communicate their understanding of a representation, such as the following graphical representation (see Figure 3).

The map and the graph below describe a car journey from Nottingham to Crawley using the M1 and M23 motorways.



Describe each stage of the journey, making use of the graph *and* the map. In particular describe and explain what is happening from A to B; B to C; C to D; D to E and E to F.

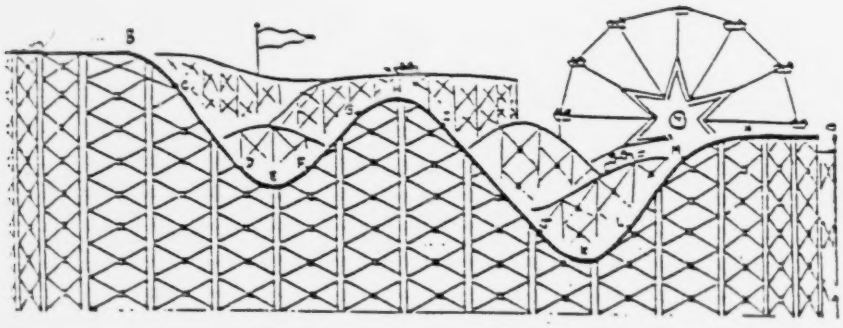
Figure 3. *Interpreting a Journey*. (Swan, 1986, p.12)

Other tasks may emphasize the level of reasoning associated with a set of questions about the same situation such as the following superitem (see Figure 4). Still other tasks may ask students to carry out a physical process, such as gather data, measure an object, construct a figure, work in a group to organize a simulation, etc. And still others may be open-ended like the following "roller coaster" problem (see Figure 5).

This is a machine that changes numbers. It adds the number you put in three times and then adds 2 more. So if you put in 4, it puts out 14.

If 4 is put out, what number was put in?
If we put in a 5, what number will the machine put out?
If we got out a 41, what number was put in?
If x is the number that comes out of the machine when the number y is put in, write down a formula that will give us the value of y whatever the value of x.

Figure 4. An *Example of a "Super Item"*. (Collis, Romberg, and Jurdak, 1986, p.12)



The picture above shows the track of a free-wheeling roller-coaster, which is travelling at a walking pace between A and B.

1. Write a paragraph describing how you think the speed of the roller-coaster varies as it travels along the track. (Use the letters A and O to help you in your description.)
2. Now sketch a graph which shows how the speed varies as it travels along the track. (Don't expect to get it right the first time.)

Figure 5. *Interpreting a Roller-Coaster*. (Swan, 1986, p.12)

These illustrations demonstrate that there are several different aspects of doing mathematics within any mathematical domain. To be able to assess the level of maturity in a domain an individual or group has achieved requires that a rich set of tasks be constructed.

> **Principle 3.** Some tasks in a particular domain would be administered to students via tailored testing (and for groups via matrix sampling as well).

Not all tasks for a domain need to be given to a student or group to determine the level of maturity. The technology is available to systematically vary several aspects of any exercise or problem situation. For example, for the subtraction exercise under Principle 1, one could vary the situations (join-separate, part-part-whole, comparison, etc.), the size of the numbers, the transformations, and the computational strategies (counting, algorithms, etc.).

> **Principle 4.** Based on the tasks administered to a student in a domain, their complexity, and the student's responses to those tasks, the information should be logically combined to yield a score for that domain.

Note that this score is not just the number of the correct answers a student has found. Instead, it would involve Boolean combinations of information (such as, following inferential rules like "if _____ and _____, then _____"). The intent of the score is that it reflect the degree of maturity the student has achieved with respect to that domain. Note that this assumes all students are capable of some knowledge in several domains.

> **Principle 5.** Construct for each individual or group a score vector over the appropriate mathematical domains. Thus, for any individual one would have several scores (X_1, X_2, ..., X_n) where X_i is the score for a particular domain.

Note that this simply reinforces the notion that mathematics is a plural noun. Rather, mathematics encompasses several related domains.

In summary, awareness of a problem, such as the need for alternative testing procedures for school mathematics, does not mean solutions are easy. It may take years to replace current testing procedures in schools. Nevertheless, this should not deter us from exploring plausible alternatives. What is needed are tasks that provide students an opportunity to reflect, organize, model, represent, argue etc. within specific domains. Constructing, scoring, scaling, and interpreting responses to such tasks for domains will not be easy, but in the long run, worth the effort.

Trends in Assessment by Teachers

One striking consequence of the scientific, psychometric assessment procedures has been to deskill teachers. External objective assessment was deemed better than professional judgment. Today, too many teachers are no longer trained in evaluation and lack confidence in their ability to judge student performance (Apple, 1979). Aware of this, a trend to empower teachers is emerging. For example, the Graded Assessment Project in England (Close & Brown, 1988) provides teachers with procedures to assess performance. This theme is central to the North American NCTM Evaluation Standards (1988). It is also a major component in the Australian MCCP project (Clarke, 1987); and is a focal part of the CGI research project at the University of Wisconsin (Peterson, Carpenter, Fennema, & Loef, 1987).

Practical-Political Trend

In most of the world, it is generally agreed that the educational system, as a whole, and the teaching of mathematics, in particular, need to change. Demands are being made of governments, politicians, and administrators for funds to bring about this reform. In turn, of course, administrators have a right to demand that evidence be gathered that their monies are well spent, that changes are made, and that the changes make a difference. Valid pupil performance data is the kind of information demanded.

However, governmental expectations about such data in the United States and Great Britain revert back to the scientific-experimental notions of the past: behavioral objectives, norm-referenced scores, Bloom's Taxonomy, ... For example, "attainment targets" in the new national curriculum in Great Britain are merely new labels for behavioral objectives. The use of SIMS items for policy profiles (e.g., in Italy and in some parts of the United States) continues the practice of not assessing problem-solving strategies, communication skills, level of reasoning, etc. These, along with other examples, make it clear that there is considerable disparity between current theory and these practical demands. The demands for information are legitimate. The validity of procedures is suspect.

Conclusions

The field of assessment and evaluation has come a long way during the last quarter century. However, a lot needs to be done. Growth in domains has been replaced with general levels of performance.

Unless changes are made in the way in which information is gathered from students, we will only contribute to the ongoing difficulties of sterile lessons, further deskilling of teachers, and so on. Instead, we

need to conceive of curricular evaluations and of assessments of individual progress in light of mathematical maturity in specific domains.

1. Current testing procedures are unlikely to provide valid information for decisions about the current reform movement.

Current tests reflect the ideas and technology of a different era and world view. They cannot assess how students think or reflect on tasks, nor can they measure interrelationships of ideas.

2. Work should be initiated (or extended) to develop new assessment procedures.

Only by having new assessment tools that reflect authentic achievement in specific mathematical domains can we provide educators with appropriate information about how students are performing. Of necessity, this implies that considerable funds be allocated for research and development. Only when new instruments are developed will we no longer be bound by old assessment procedures rooted in the traditions of the Industrial Age.

3. The emerging variety of evaluation models need to utilize assessment procedures which reflect the changes in school mathematics.

Today, school mathematics is changing the emphasis from drill on basic mathematical concepts and skills to explorations that teach students to think critically, to reason, to solve problems. The criteria for judging level of performance by a student or group of students should be based on these notions. This will involve the student's capability—when posed with a problem situation in a specific mathematical domain—of communicating, reasoning, modeling, solving, and verifying propositions. Also, the index or scale developed to measure performance should reflect the student's level of maturity in that domain.

References

Apple, M. W. (1979). *Ideology and curriculum*. London: Routledge & Kegan Paul.

Ayres, L. P. (1918). History and present status of educational measurements. In S. C. Parker (Ed.), *The measurement of educational products: Seventeenth yearbook of the National Society for the Study of Education* (Part II). Bloomington, IL: Public School Publishing Co.

Begle, E. G., & Wilson, J. W. (1970). Evaluation of Mathematics Programs. In Begle, E.G.(ed.), *Mathematics Education, 69th Yearbook of the NSSE (Part 1)*. Chicago: University of Chicago Press.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook I: Cognitive domain. New York: Longmans, Green, & Company.

Campbell, D. T., & Stanley, J. T. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Clarke, D. (1987). The interactive monitoring of children's learning of mathematics. *For the Learning of Mathematics*, 7(1), 2-6.

Close, G., & Brown, M. (1988). *Graduated assessment in mathematics: Report of the SSCC study*. London: Department of Education and Science.

Collis, K. F. (1987). Levels of reasoning and the assessment of mathematical performance. In T. A. Romberg (Ed.), *The monitoring of school mathematics: Background*. Madison, WI: Wisconsin Center for Education Research.

Collis, K. F., Romberg, T. A., & Jurdak, M. E. (1986). A technique for assessing mathematical problem-solving ability. *Journal for Research in Mathematics Education*, 17(3), 106-221.

de Lange, J. (1987). *Mathematics, insight, and meaning*. Utrecht, Holland: Rijksuniversiteit Utrecht.

Eash, M. J. (1985). Evaluation research and program evaluation: Retrospect and Prospect: A reformulation of the role of the evaluator. *Educational Evaluation and Policy Analysis*, 7(3). 249-252.

Eisner, E. Educational Connoisseurship and criticism: Their form and function in educational evaluation. *Journal of Aesthetic Education*, 10, 173-179.

Fetterman, D. (1984) *Ethnography in educational evaluation*. Beverly Hills: Sage.

Foxman, D. D., Cresswell, M. J., Ward, M., Badger, M. E., Tucson, J. A., & Bloomfield, B.A. (1980). *Mathematical development primary survey report no. 1*. London: Her Majesty's Stationery Office.

Foxman, D. D., Badger, M. E., Martini, R. M., & Mitchell, P. (1981). *Mathematical development secondary survey report no. 2*. London: Her Majesty's Stationary Office.

Freeman, F. N. (1930). *Mental tests: Their history, principles and applications* (rev. ed.). Boston: Houghton Mifflin.

Gorth, W. P., Schriber, P. E., & O'Reilly, R. P. (1974). *Comprehensive achievement monitoring: A criterion-referenced evaluation system.* New York: Educational Technology Publishers.

Greene, H. A., Jorgensen, A. N., & Gerberich, J. R. (1953). *Measurement and evaluation in the elementary school.* (2nd ed.). New York: Longmans.

Guba, E., & Lincoln, Y. (1981) *Effective evaluation.* San Francisco: Jossey Bass.

Guralnik, P. B. (Ed.). (1985). *Webster's New World Dictionary.* New York: Prentice Hall.

McLean, L. D. (1982). *Report of the 1981 field trials in English and mathematics: Intermediate division.* Toronto, Ontario: The Minister of Education.

National Coalition of Advocates for Students. (1985). *Barriers to excellence: Our children at risk.* Washington, D. C.: Author.

National Council of Teachers of Mathematics. (1988). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Odell, C. W. (1930). *Educational measurements in high school.* New York: Century.

O'Keefe, J. (1984) The impact of evaluation on federal education program policies. *Studies in Educational Evaluation,* 10, 61-74.

Patton, M. Q. (1980) *Qualitative evaluation methods.* Beverly Hills: Sage.

Peterson, P. L., Fennema, E., Carpenter, T. P., & Loef, M. (1987). Teachers' pedagogical content beliefs in mathematics. *Cognition and Instruction.*

Popkewitz, T. S. (1984). *Paradigm & ideology in educational research.* London: The Falmer Press.

Reinhard, D. (1972) Methodology for input evaluation utilizing advocate and design teams. Unpublished PhD. dissertation. The Ohio State University.

Romberg, T. A. (1987). The domain knowledge strategy for mathematical assessment. Project Paper #1. University of Wisconsin.

Romberg, T. A. (Ed.). (1985). *Toward effective schooling.* New York: University Press of America.

Romberg, T. A. (1983). A common curriculum for mathematics. In G. D. Fenstermacher, & J. J. Goodlad (Eds.), *Individual differences and the common curriculum.* Chicago: The University of Chicago Press.

Romberg, T. A. (1976). *Individually guided mathematics.* Reading, MA: Addison-Wesley.

Romberg, T. A. (1975). Answering the question—is "it" any good?—The role of evaluation in multi-cultural education through competency-based teacher education. In C. A. Grant (Ed.), *Sifting and winnowing: An exploration of the relationship between multi- cultural education and CBTE.* Madison, WI: Teacher Corps Associates.

Schoenfeld, A. H., & Herrmann, D. J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology:* Learning, Memory, and Cognition, 8, 484-494.

Scriven, M. (1974). Evaluation perspectives and procedures. In W. J. Popham (Ed.), *Evaluation in Education.* Berkeley: McCutchan.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology,* 15, 201-293.

Stake, R. E. (1974). Program Evaluation, particularly responsive evaluation. Occasional Paper No 5. Calamus: Western Michigan University Evaluation Center.

Stake, R. E., & Gjerde, C. (1974). An evaluation of the T-CITY. In Kraft et al. (Eds.) *Four evaluation examples: Anthropological, economic, narrative and protrayal.* AERA Monograph Series on curriculum evaluation, No. 7. Chicago: Rand McNally.

Swan, M. (1987). *The language of functions and graphs.* Nottingham: Shell Centre for Mathematical Education.

Swan, M. (1986). *The language of graphs: A collection of teaching materials.* Nottingham, England: The Shell Centre for Mathematical Education.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements.* New York: Teachers College, Columbia University.

Tyler, R. W. (1931). A generalized technique for constructing achievement tests. *Educational Research Bulletin,* 8, 199-208.

Vergnaud, G. (1982). Cognitive and developmental psychology and research in mathematics education: Some theoretical and methodological issues. *For the Learning of Mathematics,* 3(2), 31-41.

Watson, G. (1938). The specific techniques of investigation: Testing intelligence, aptitudes, and personality. In G. M. Whipple (Ed.), *The scientific movement in education: Thirty- seventh yearbook of the National Society for the Study of Education* (Part II, pp. 365-366). Bloomington, IL: Public School Publishing.

Weinzweig, A. I., & Wilson, J. W (1977). Second IEA mathematics study: Suggested tables of specifications for the IEA mathematics tests. *Working Paper I*. Wellington, New Zealand: IEA International Mathematics Committee.

Young, M. F. D. (1975). An approach to the study of curricula as socially organized knowledge. In M. Golby, J. Greenwald, & R. West (Eds.), *Curriculum design* (pp. 101-127). London: The Open University Press.

Part 2

Papers *from* *the Second International Mathematics Study*

CURRICULUM-LINKED ASSESSMENT: A MODEL BASED ON THE SECOND INTERNATIONAL MATHEMATICS STUDY

Kenneth J. Travers

The Second International Mathematics Study (SIMS) was a comprehensive survey of the teaching and learning of mathematics in the schools of some twenty countries (educational systems) around the world. The Study was conducted under the aegis of the International Association for the Evaluation of Educational Achievement. In the Study, detailed information was obtained on the content of the implemented mathematics curriculum, what mathematics was actually taught by the teachers, and how that mathematics was taught. Student achievement and attitudes were assessed using internationally developed tests and questionnaires that were taken by random samples of mathematics classes in each country. The Study was targeted at 13-year olds in most countries (12-year olds in Japan and Hong Kong), and at those students at the end of secondary school who were enrolled in advanced college-preparatory mathematics courses. The lower level, younger group was called "Population A"; the older group, "Population B".

For each target population, topics were tested that reflected an international consensus of mathematical content judged to be important by panels of experts in each country. As a result, the fit of the tests to the curriculum varied somewhat from country to country. Data were obtained from teachers as to whether the content had been taught to the students who were tested. This information, called "opportunity-to-learn," provided a backdrop for interpreting the achievement scores. In each participating country, SIMS was carried out by a nationally recognized educational research institution under the direction of a national committee of specialists in mathematics education and educational research.

The First International Mathematics Study took place in 1964 in twelve countries. Eleven of these countries, including the United States and Japan, participated in the second study in 1980-82.

In the United States, students in approximately 500 mathematics classrooms in about 250 public and private schools randomly selected from across the country were tested at the end of the 1981-1982 school year. (A number of countries, including the United States, also tested the students at the beginning of the school year.) The countries (systems) taking part in the Study were:

Belgium (French & Flemish)	Luxembourg
Canada (British Columbia & Ontario)	Netherlands
England and Wales	New Zealand
Finland	Nigeria
France	Scotland
Hong Kong	Swaziland
Hungary	Sweden
Israel	Thailand
Japan	United States

The SIMS Model

The Second International Mathematics Study was based on three aspects of the curriculum: the intended curriculum, the implemented curriculum, and the attained curriculum. The intended curriculum is reflected in curriculum guides, course outlines, syllabi, and textbooks adopted by school systems. In most countries, national curricula emanate from a ministry of education or similar body. In the United States such statements of intended goals and curricular specifications come from the Department of Education in each state and from local districts. Thus, it was considerably more difficult to describe the intended curriculum for the United States than for almost any other country that took part in the study.

The implemented curriculum focuses on the classroom, where the teacher interprets and puts into practice the intended curriculum. Teachers exercise their own judgment in translating curriculum guides and adopted textbooks into programs for their classes. Hence, their selection of topics or patterns for emphasis may not be consistent with those intended.

To identify the implemented curriculum, a number of questionnaires were developed for classroom teachers to complete. For example, teachers were asked whether or not they had provided instruction for each of the items on the achievement tests. They were questioned about such matters as the use of calculators in their classes. They were also asked to provide detailed information on the number of class periods that they devoted to specific topics and subtopics and on how they presented and interpreted this mathematical content to their classes.

The attained curriculum is what students have learned as measured by tests and questionnaires. Exten-

sive achievement tests were designed to assess student knowledge and skills in areas of mathematics designated as important and appropriate for the students being tested. The fit between these tests and the actual curricula in individual countries varied considerably. The tests contained items that were less appropriate in some countries than in others. Furthermore, the tests could not possibly contain an adequate range of items to fully represent all curricula in all countries.

The student outcome measures also included a number of opinion surveys and attitude scales. These were devised to elicit student views on the nature, importance, ease, and appeal of mathematics in general and of selected mathematical processes.

The SIMS model provides an array of background information for viewing student outcomes. That is, one can regard cross-national patterns of achievement in the light of the content of the (intended) curriculum in each country and teacher coverage (opportunity-to-learn) of that content. Therefore, the model enables a triangulation on student outcomes. For some countries, two additional sources of data were available. Those data points are (i) pre-test data — students were tested at the beginning of the school year, as well and (ii) classroom process data — detailed information on how the teacher handled the subject matter as it was presented during the school year.

SIMS as a Model for Assessment

The SIMS model lends itself to powerful approaches to program (curriculum) assessment. Note, for example, Cronbach's (1964) distinction between every-pupil testing and evaluation for course improvement. Cronbach has noted that the concern in every-pupil testing is for precise and valid comparisons among individuals (for purposes, say, of making decisions about promotion, selection or reporting). As Cronbach has noted:

> Much of test theory and test technology has been concerned with making measurements precise. Important though such precision is for most decisions about individuals, I shall argue that in evaluating courses we need not struggle to obtain precise scores for individuals...(p. 233)

SIMS, as an activity in program assessment (as contrasted with testing for making decisions about individual students), has the following features.

1. Curriculum Coverage — item sampling

Since the interest in program assessment is not in scores for individual students, but in how well a body of subject matter has been learned by a cohort of students, SIMS used an item-sampling scheme for testing. Under this plan, a comprehensive set of mathematics items was responded to by the entire class. However, within the class, subsets of items were answered by a fraction of the students. (The eighth grade test has only 180 items and the twelfth grade test has 136 items.)

2. Test Scores

In program evaluation, interest resides not in a single test score, but in achievement at the subscore and item level. As Harnisch and Linn (1981) point out, a score of 10 on a 20 item test could have been arrived at in 184,756 ways. Again, Cronbach (1964) states:

> Outcomes of instruction are multidimensional, and a satisfactory investigation will map out the effects of the course along these dimensions separately ... To agglomerate many types of post-course performance into a single score is a mistake, since failure to achieve one objective is masked by success in another direction. Moreover, since a composite score embodies (and usually conceals) judgements about the importance of the various outcomes, only a report that treats the outcomes separately can be useful to educators who have different value hierarchies. (page 236)

3. Curriculum-linked vs. curriculum-free testing

Much of large scale testing in the United States entails tests of general intellectual development or aptitude that are often used as criteria for school achievement or effectiveness. SIMS, by contrast, focuses on the mathematical content of the curriculum, as found in the syllabus or textbook, as taught by the teacher and as learned by the student.

As Madaus (1979) has stated:

> It has been argued that although tests of general intellectual development or intelligence do not measure the behavioral objectives of specific programs, they are in fact the best criteria we have of general educational development (Cooley and Lohnes, 1976). This may be so, but it seems odd to measure what is admittedly a side effect of education while at the same time ignoring the more direct results of particular curricula and courses. ...Conclusions about the direct instructional effects of schools should not have to rely on evidence relating to skills taught incidentally. (Madaus, et al., 1979).

Curriculum Analysis within the SIMS Framework

SIMS was developed on the basis of a survey of the mathematics curriculum for each target population in each participating country. Consequently, informa-

tion about the curriculum is a fundamental product of SIMS. A framework for studying the curriculum, and in developing the international item pools, was a grid that consists of rows for mathematical topics and columns for behavioral levels at which the topics are considered.

The design of SIMS facilitates a detailed analysis of curriculum for a country or educational system. Activities that may be undertaken at the system level in order to assemble information that is useful to have in order to better understand SIMS data include:

a. A detailed look at the system's mathematics curriculum, from the perspective of the SIMS grid and item pool. Information is provided not only on the content dimension (what subject matter is in the curriculum), but on the behavioral dimension (what cognitive levels are perceived to be emphasized in the curriculum).

b. Identification of curricular areas that are emphasized or not emphasized (again with respect to SIMS).

The above information is most useful in helping to interpret data from the teachers on opportunity to learn (e.g. To what extent do teachers cover vectors, an important topic in the curriculum? and student achievement (e.g. Can low student achievement in probability be attributed to low teacher coverage?)

Why Replicate SIMS?

SIMS provides a mechanism whereby an educational system (a state, a province, or a school district) can obtain detailed information on its curriculum as intended, implemented and achieved. The exercise of analyzing the content of the intended curriculum can serve to identify, within a common framework, those aspects of mathematics that are emphasized and those that are of less importance. With such data in hand, curriculum supervisors are then able to make more informed decisions about the system's goals for mathematics education. Since the SIMS framework is international in scope, educational personnel have the opportunity to make comparisons not only with their own national system, but with those of other countries.

At the level of the implemented curriculum, data on teacher coverage of various topics can be useful, for example, as a basis for designing in-service programs. Consider a system where achievement in probability and statistics is found to be much lower than desired. Assume further that it is found that teacher coverage of these topics is low. It may be that the topics are in Chapter 15, at the end of the textbook, and teachers tend to not get that far. Or it may be that teachers avoid the topic since they feel unprepared to teach it. Programs of professional development could then be designed to assist the teachers in greater managing instructional time to enable better coverage of important topics. Alternatively, workshops could be devised to upgrade teachers' subject matter and instructional competencies.

Another use of a SIMS replication may be that of a tool for assisting in implementing a new curriculum. In this time of curricular change a variety of frameworks are being proposed for revising the instructional program for a school or district. The SIMS model may provide "benchmarks" for use in assessing the degree to which curriculum reform has occurred over a period of time. For example:

Intended Curriculum. The SIMS curriculum analysis can help identify aspects of a system's curriculum that are aligned with the desired plan (framework) as well as those aspects still needing refinement.

Implemented Curriculum. An analysis of the data on teacher coverage (opportunity to learn) can help identify topics and strategies that need further attention (say, through in-service programs).

Summary

The Second International Mathematics Study is based on a model that views the curriculum as intended (e.g. content of syllabi, courses of study), as implemented (content actually taught by the teacher), and as attained (mathematics learned by the student). Consequently, patterns of achievement (either within or between educational systems) may be examined against a background of detailed information on the content of the curriculum both as intended to be taught and as actually (reported to be) taught. Such detailed curricular data may be useful to curriculum supervisors and evaluators, for example, as they assess present curricula, plan new programs and seek to document the extent to which curricular innovation has taken place.

The kinds of data which may be obtained from SIMS replications within countries include:

a. Background data of a great variety: e.g., characteristics of schools, teachers and students.

b. Curricular content data: e.g. what topics are in the curriculum for each target population, in the various countries.

c. Teacher coverage data: e.g. between countries: What topics receive what level of coverage? Is algebra taught to junior high school age students (12-13 yrs.) in all SIMS countries? Within countries: e.g. Are all students, or only those in upper academic tracks, taught algebra?

References

Burstein, L. (1976). The Choice of unit of analysis in the investigation of school effects: IEA in New Zealand. *New Zealand Journal of Educational Studies*. 11, 11-14.

Cronbach, L.J. (1964). Evaluation for course improvement. In R.W. Heath, (Ed.) *New Curricula*. New York: Harper and Row, pages 231-248.

Harnisch, D.L. & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*. Vol. 18, No. 3, pages 133-146.

Lewy, A. (1973). Discrimination among individuals v. discrimination among groups. *Journal of Educational Measurement*. 10, 19-24.

Madaus, George, et al. (1979). The sensitivity of measures of school effectiveness. *Harvard Educational Review*, 49 pages.

Travers, K.J. & C.C. McKnight (1985). Mathematics achievement in U.S. schools: Preliminary findings from the Second International Mathematics Study. *Phi Delta Kappan*, pp. 407-413.

CALCULUS IN THE HIGH SCHOOLS OF THE UNITED STATES OF AMERICA AND CANADA (ONTARIO)

Michael K. Dirks · David F. Robitaille · John Leduc

The teaching of calculus at the secondary school level in North America has been and remains a controversial matter. On the one hand, some college and university mathematicians argue that this course belongs exclusively within their jurisdiction (Henry, Jones, and Kenelly, 1985). They may question the equivalence of the high school course with their own as well as bemoan the lack of basic preparation which incoming high school students have in pre-calculus topics. On the other hand, some educators assert that in countries such as Japan and the Soviet Union a much greater proportion of the age cohort successfully studies calculus at the secondary level, and imply that this fact enhances the ability of these countries to compete industrially or militarily (Wirszup, 1980). In this report data collected in 1981-82 as part of the Second International Mathematics Study are used to describe the teaching and the learning of calculus at the secondary level in the United States and in the Canadian province of Ontario. A number of achievement comparisons with other jurisdictions are also included to provide a better basis for drawing conclusions.

The Teaching of Calculus

Approximately 125 Grade 13 classes from Ontario and 175 Grade 12 classes from the United States participated in the Second International Mathematics Study at the senior secondary level. Not all of these classes studied calculus, however, and the present report is based on responses from the 62 Canadian and 44 American classes who were in fact taking a calculus course, and whose teachers completed the Calculus Questionnaire.

Of the 44 American classes, 43 reported spending a full year on calculus and the other spent more than one semester but less than a year on the subject. Of the 62 classes in Ontario, 51 reported spending a full school year on calculus, 10 spent more than a semester but less than a year, and one class studied one chapter of calculus. Twenty six of the teachers in the United States reported that their students took the Advanced Placement (AP) examination, with 23 taking the AB exam and 3 taking the BC exam. None of the Canadian schools indicated that their students had taken the AP examination.

Curriculum Materials and Course Content

The two groups differ markedly in the textbooks they used, as well as in the amount of supplementary materials which teachers reported producing for their classes. The Canadian teachers used six different texts with two of these texts accounting for 55 classrooms. At least a dozen different books were used in the United States and only one was used in more than five classrooms. Of all the different texts in use, only two were used in calculus classrooms in both countries and only in five classrooms. The majority of the Canadian teachers reported using a text which provided a "somewhat intuitive treatment of calculus" and which "might be described as 'pre-college' calculus texts" (Alexander, 1987). Most of the American teachers reported using a standard, college-level calculus text.

While only a minority of teachers supplemented the text with materials which they developed themselves, those in Ontario did so more frequently than did their American counterparts. For example 14 of the 36 Canadian teachers who taught integration by trigonometric substitution indicated that they had developed supplementary materials while only 2 of 32 American teachers reports developing such materials for this topic.

The Calculus Questionnaire was designed to obtain information on the teaching of 21 topics. Teachers were asked whether or not they taught the topic, how it was taught (as new material, reviewed, and extended, or assumed as prior knowledge), how difficult the topic was to teach and to learn, what influenced their decision to teach a topic, and whether or not the topic was in the text.

Responses from teachers in Ontario and from the United States indicated that 10 of the 21 topics were almost universally taught in both jurisdictions: limit of a function; limit as x approaches infinity; derivative of a polynomial function; derivative of a sum, a difference, a product, and a quotient of functions; chain rule for differentiation; implicit differentiation; related rates; relative extrema; definite integral as the area under a curve; and calculus of exponential and logarithmic functions. These topics were taught in most classes as new material, with a few of the American classes treating the topics as material to be reviewed and extended.

The percent of teachers who included each of the 21 topics in their courses, or who assumed that a topic had been previously learned, are displayed in Table 1. Since teachers very rarely indicated that a topic was assumed as prior knowledge or reviewed, this information is not shown separately. Table 1 also includes the percent of teachers who responded that a topic was in

the student text. The results indicate that American students were more likely to cover more of the typical first year college calculus topics than were students in Ontario. This is particularly true for the topics of continuity, arc length, methods of integration, and indeterminate forms.

Table 1
Frequency with Which Topics Appear in Classes and Texts

Topic	Taught or assumed as prerequisite (Percent)		Present in text (Percent)	
	USA	Ontario	USA	Ontario
Limit of a sequence	73	95	68	70
Limit of a function	100	97	98	69
Limit as x approaches infinity	98	98	96	66
Continuity	100	56	93	50
Derivative of a polynomial function	100	100	98	81
Derivative of a sum, difference product or quotient of functions	100	98	98	78
Chain rule for differentiation	100	98	96	81
Implicit differentiation	100	98	96	81
Related rates problems	95	98	93	79
Relative extrema	98	92	96	73
Definite integral as area	100	94	98	90
Arc length	67	34	82	35
Calculus of logarithmic and exponential functions	98	92	96	81
Indeterminate forms and l'Hopital's rule	66	18	75	14
Integration by trignometric substitution	74	57	93	41
Integration by parts	82	54	82	47
Integration by partial fractions	57	57	80	46
Numerical integration	64	25	80	27
Series	24	57	48	43
Partial derivatives	7	10	46	9
Multiple integrals	2	8	39	9

Data were also gathered on the number of class periods spent on each of the twenty-one topics, and results for the twelve most frequently taught topics are displayed in Figure 1.

The boxplots indicate that American teachers tended to devote more periods to most of these topics than their Canadian counterparts.

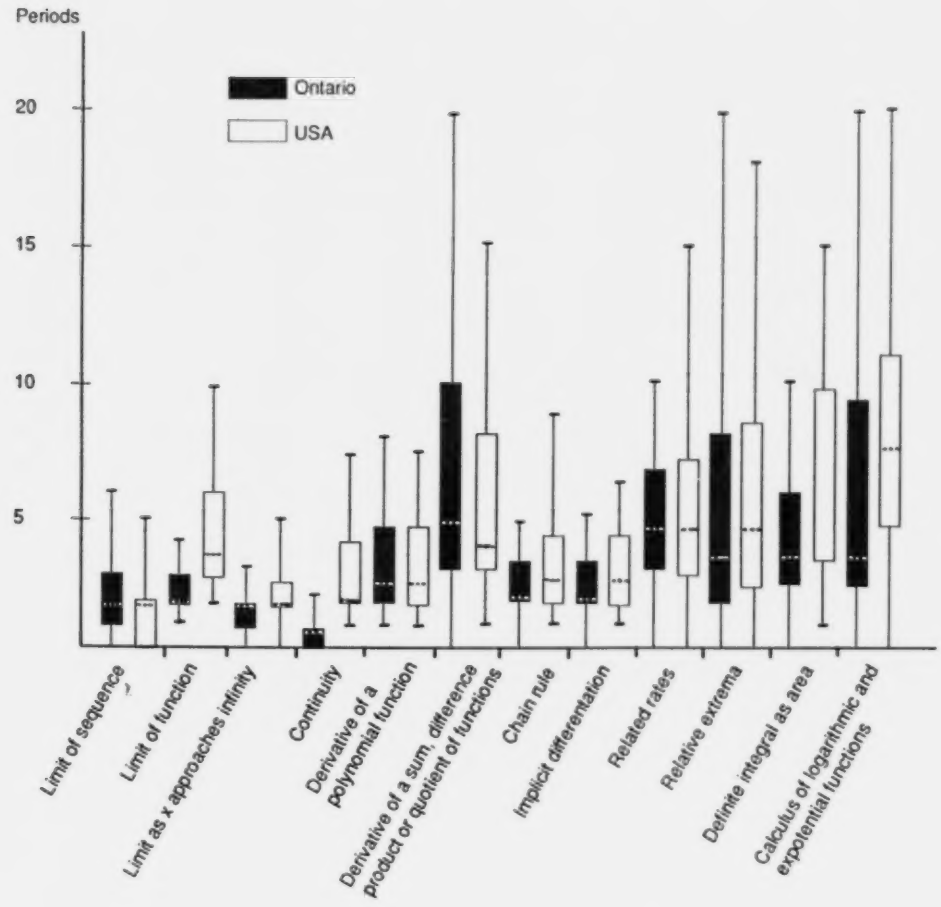


Figure 1. Duration of presentations for selected topics.

Classroom Activities, Instructional Aids, and Applications

Teachers were asked to estimate the percent of time within a typical period that was spent on each of three types of activities: teacher presentation or teacher led review, teacher and student discussion of homework, and student activities and supervised study. The estimates were remarkably similar with medians for both the United States and Ontario near 40, 30, and 20 percent respectively. Canadian teachers tended to use slightly more time for presentation and review and slightly less time on homework.

Few teachers of calculus appear to make use of teaching aids other than the textbook, the overhead projector, and the hand-held calculator. In particular, only one-fifth of the American teachers said they made use of movies in their courses while almost no Canadian teachers did so. Twenty-five percent of Americans reported using physical models compared with 15 percent of the Canadians. In contrast, calculators were used in almost every Canadian classroom and in 80 percent of American ones. Computers or micro-computers were used in only about ten percent of the classes surveyed. This figure, one would hope, may have risen sharply since these data were collected.

Calculus has applications in many fields of study in the physical, biological, and social sciences. However, these teachers reported that the vast majority of the time they devoted to applications in their calculus classes was in the areas of applications in the fields of physics and engineering. Applications from business ranked a distant third, as is shown in the boxplots in Figure 2.

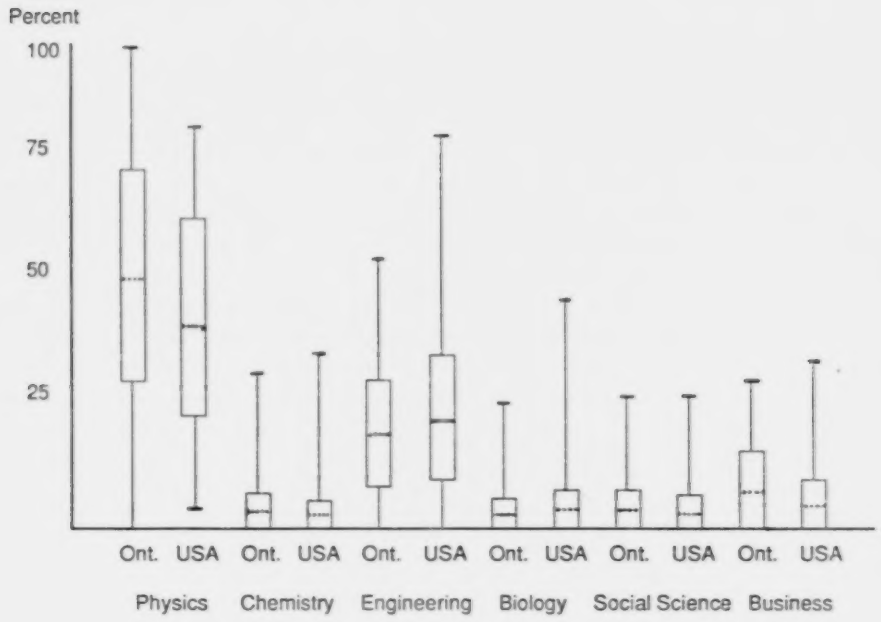


Figure 2. Percent of time spent on applications from various fields.

Applications from chemistry, biology, and the social sciences were almost never considered. Given the importance that calculus now has in these areas and the number of students who will study them in college, this would seem to be an unfortunate state of affairs. Teachers, curriculum developers, and textbook publishers should be aware of the need to broaden their coverage of the applications of calculus, and calculus courses and textbooks should include many more examples of applications from other areas.

Teachers were also asked about sources of applications. Most teachers indicated that the applications they presented were drawn from both the textbook in use and from supplementary textbooks. In addition, 34 percent of the Americans and 60 percent of the Canadians reported creating applied problems themselves. Other sources, such as professional journals and meetings, were seldom mentioned. Teachers were asked specifically if they utilized the UMAP application modules, but none reported doing so. It must be added, however, that implementation of the UMAP materials was directed at the university, and not at the high school level.

Content-Specific Teaching Methods

One of the unique aspects of the classroom process questionnaires developed for use in the Second International Mathematics Study was the collecting of data related to the methods used by teachers to teach specific concepts and skills. The Calculus Questionnaire explored how teachers handled some of the basic formulas, concepts, and theorems of calculus.

Formulas: A number of formulas involving the derivative and the integral are usually developed in a first course in calculus. Each formula might be developed through a formal proof, in informal derivation, or it might be stated without any derivation or justification. In turn, teachers' expectations for students might also vary. Teachers were asked about the teaching of six such formulas, and their responses are summarized in Figure 3.

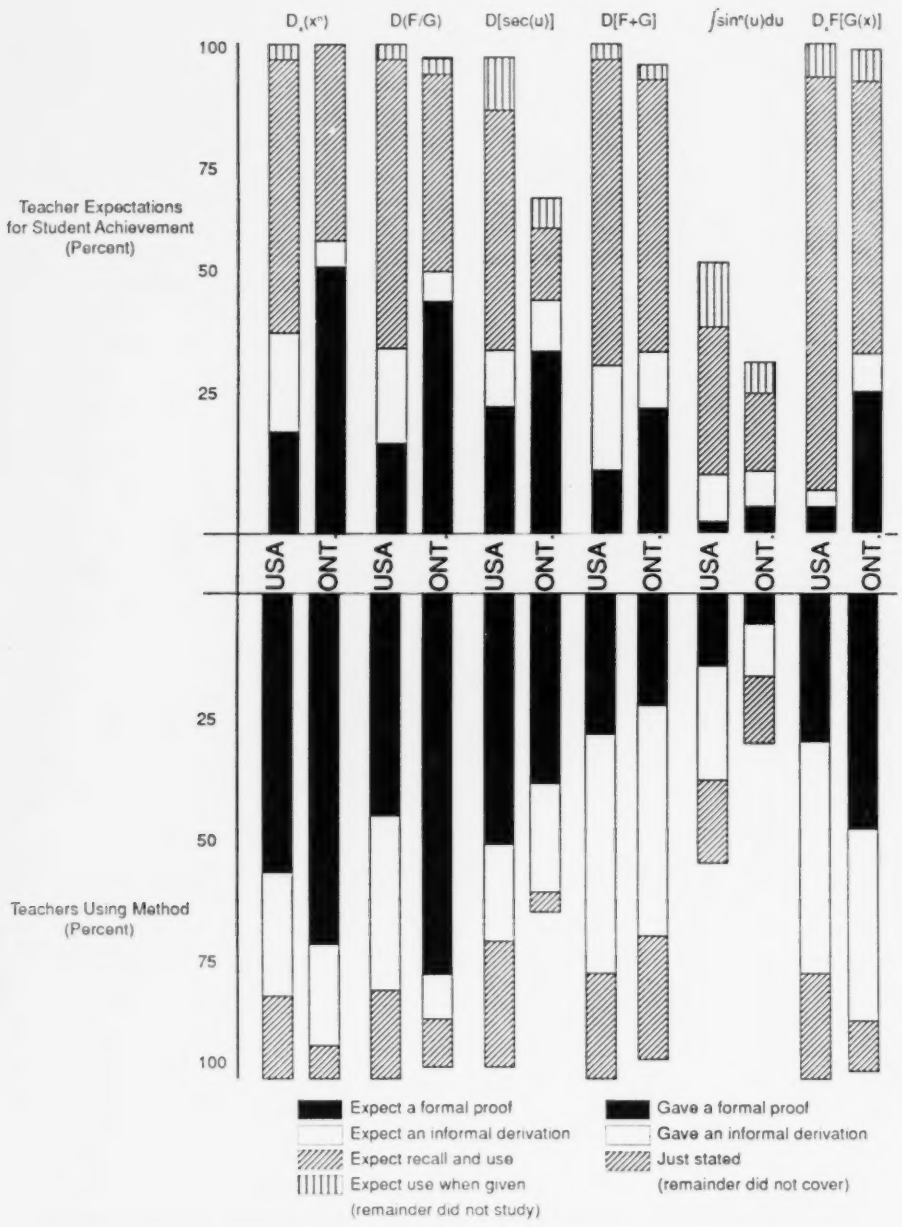


Figure 3. Mode of presentation and level of expectation for various formulas.

Taken at face value it would seem that the Canadian teachers are much more formal in their presentations and expectations for these topics. Differences are especially large for student expectations. For example, slightly more Canadian teachers than American teachers gave a formal or informal derivation of the quotient rule for derivatives: 89 percent compared with 82 percent. Many more Canadian teachers, however, required that their students also be able to provide a justification of this result: 53 percent compared with 33 percent. Similarly, although Canadian teachers were only slightly more likely to provide a derivation of the chain rule for their students, they were four times as likely to expect that their students would also be able to justify this result. It can be argued that without holding students responsible for justifying the formulas which they learn at some level, as the Canadians are more apt to do, these formulas will be less meaningful and more easily forgotten.

Questions were also asked relating to the teaching of four basic concepts: limits, continuity, the derivative, and the integral. The questions and responses are given below,and when these responses are compared to those in Figure 3 some interesting questions arise.

Limits and Continuity: To determine which of several approaches were used in the teaching of the concept lim $f(x) = L$ as x approaches a, teachers were asked to indicate the methods which they used from several choices. The statement and choices that they were given and the frequency of responses are listed below in the box.

Several observations can be based on these responses. First, the frequency with which teachers reported using graphical arguments to support the concept of the limit of a function at a particular point is surprisingly low. It might be expected that graphs would be universally employed but barely half of the Ontario teachers reported using them. The American teachers used graphical arguments more often, but over a third did not do so.

A second observation has to do with the epsilon-delta and deleted neighborhood approaches to limits. Both of these approaches were used much more frequently in American classrooms. A formal epsilon-delta definition was given by over 80 percent of the American teachers but by less than 20 percent of their Canadian counterparts. This is probably a reflection of the college texts usually used in the United States. Since the AB version of the Advanced Placement test does not include questions on epsilonics, preparation for this test would not in and of itself require teachers to use this formal approach to limits. Whether or not students' early encounters with the limit concept should involve epsilonics is a matter for debate. It would seem clear, however, that if such a formal approach is employed that it should be accompanied by a graphical interpretation. From the responses to this question, however, it appears that a sizeable number of teachers use delta-epsilon definitions without supporting graphs.

As with the questions on formulas, teachers were asked what was expected by way of student concept attainment. Specifically, they were asked what they expected their students to be able to do after the concept lim $f(x)=L$ as x approaches a had been taught. The statement and choices that they were given and the frequency of responses are listed in the box on the next page.

In the teaching of the concept of $\lim_{x \to a} f(x) = L$

(a) I discuss how as x "gets close to a," f(x) "gets close to L."
Ontario 87% USA 82%

(b) I use the formal definition of epsilon and delta.
Ontario 18% USA 82%

(c) I use the concept of limit of a sequence.
Ontario 62% USA 21%

(d) I use the concept of elements in a deleted neighborhood of a being mapped into a neighborhood of L.
Ontario 3% USA 48%

(e) I develop it intuitively with graphical arguments involving the graphs of particular functions.
Ontario 52% USA 64%

(f) I did not discuss limits with the target class.
Ontario 3% USA 0%

After teaching the concept of $\lim_{x \to a} f(x) = L$ I expect my students to be able to:

(a) evaluate the limit of a first degree polynomial function.
Ontario 94% USA 89%

(b) state the epsilon-delta definition of the limit of a function.
Ontario 13% USA 71%

(c) give an epsilon-delta proof that the limit of $f(x) = 2x + 5$ is 9 as $x \to 2$
Ontario 6% USA 77%

(d) give an epsilon-delta proof that the limit of $f(x) = \frac{1}{x^2}$ is $\frac{1}{4}$ as $x \to 2$
Ontario 3% USA 52%

(e) use the epsilon-delta definition of a limit of a function to prove that if $\lim_{x \to a} f(x) = L$ and $\lim_{x \to a} g(x) = M$, then $\lim_{x \to a} f(x) + g(x) = L + M$
Ontario 2% USA 23%

(f) I did not discuss limits with the target class.
Ontario 3% USA 0%

Since Canadian teachers seldom used epsilonics, it is not surprising that they seldom expected their students to do so. The rather high level of student expectation in the United States is somewhat surprising, however, considering the much lower expectations observed above for justifying formulas. Over 70 percent of the American teachers expected their students to be able to state an epsilon-delta definition and to employ it to justify limits of linear epsilon-delta proofs for limits of simple rational functions. The expectation that students could prove that the limit of the sum of two functions equals the sum of the limits was much lower. Just over 20 percent of the American teachers expected student proficiency for this task.

Teachers were also asked if they presented a formal definition for the concept of continuity of a function. The responses elicited are consistent with those for the limit of a function. Teachers in the United States were far more likely to use a formal epsilon-delta approach than those in Ontario. Almost 70 percent of the American teachers used a formal definition to prove that functions were continuous at specific points compared to just over 20 percent of the Canadians

Teachers were then asked what they expected their students to be able to do after the concept of continuity had been taught. While almost 70 percent of American teachers expected their students to be able to state and apply a formal definition of continuity, the corresponding figure in Ontario was only 14 percent.

Clearly, teachers in the two jurisdictions held differing views on the importance of continuity. At the most basic level, nearly all of the American teachers expected their students to be able to identify graphs of continuous and discontinuous functions compared with just half of the Canadians.

Derivative: Teachers were asked how they introduced the derivative of a function $f(x)$ at $x=a$. The differences so notable in the approaches to the first two concepts practically disappeared in this section. However, the similarity may be misleading given the underlying differences in approach to limit and continuity. The statement, choices and frequency of responses are listed at the top of the next page.

When introducing the derivative of a function $f(x)$ at $x=a$, I discuss:

(a) the rate of change of $y(f(x))$ with respect to x in the function.
Ontario 73% USA 73%

(b) the limiting position of a secant line connecting the points $(x,f(x))$ and $(a,f(a))$, as x approaches a, on the curve $y=f(x)$.
Ontario 94% USA 86%

(c) $\lim_{\Delta x \to 0} (\frac{\Delta y}{\Delta x})$
Ontario 87% USA 82%

(d) and use a formal definition of the derivative such as
$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$
Ontario 95% USA 98%

(e) I did not discuss derivatives with the target class.
Ontario 0% USA 0%

While a formal definition was almost universally given for the derivative, over a quarter of the teachers in ' oth jurisdictions reported that they did not interpret the derivative as a rate of change. Considering the importance of this notion in applying the derivative, such an omission seems quite odd.

Teachers' expectations for their students are pictured through the following responses:

Consider the following form of the definition of the derivative of a function f at a point a:
$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

(a) I do not present this definition, or any equivalent form, to my students in the target class.
Ontario 0% USA 2%

(b) I present this definition, but I do not expect the students to either remember it or use it.
Ontario 3% USA 2%

(c) I present the definition and I expect the students to be able to use it in deriving general results about derivatives of specific functions, such as finding $f'(2)$ for
$$f(x) = x^2 - x + 4$$
Ontario 86% USA 96%

(d) I present the definition and expect the students to be able to use it in deriving general results about derivatives.
Ontario 71% USA 64%

(e) I present the definition and expect the students to be able to use it in testing functions, such as $y=x$, for the existence of derivatives at points such as $x=0$.
Ontario 14% USA 50%

These responses indicate that teachers in both places expect that students will be able to state and apply the definition of derivative. Canadian teachers apparently expect their students to be able to use the definition to derive such results as the sum and product somewhat more frequently than the American teachers who were much more apt to expect their students to use the definition to test specific functions for derivatives at particular points.

Integration: The last major concept covered in the questionnaire was the definite integral. The questions were limited to an outline of a definition of this concept. The statement, choices and frequency of responses are given below:

Consider a formal definition of the definite integral as the limit of a Riemannian sum which might might involve:

$$\int_a^b f(x)dx = \lim_{n\to\infty}\sum_{k=o}^{n}(f(c_k))(x_{k+1} - x_k)$$

(a) I do not present this definition of the definite integral.
Ontario 51% USA 16%

(b) I present this definition but do not use it.
Ontario 14% USA 18%

(c) I did not discuss any interpretation of the definite integral with the target class.
Ontario 0% USA 0%

(d) I present this definition and use it to find the value of certain definite integrals, e.g.,

$$\int_0^1 x^2\,dx$$

Ontario 25% USA 61%

(e) I present this definition and use it to present some general theorems about the definite integral.
Ontario 2% USA 2%

(f) I present this definition and then immediately drop it in favor of specific rules for evaluating definite integrals.
Ontario 8% USA 2%

(g) I did not discuss the definite integral with the target class.

Ontario 0% USA 0%

The responses indicate that while teachers in both jurisdictions discuss and interpret the definite integral, only half of the Canadians use a formal definition based on the idea of a Riemann sum. The Canadian teachers either teach the definite integral very informally or they use a different definition. The majority of American teachers employ this definition to evaluate specific integrals. Teachers were asked in another questionnaire item if they interpreted the definite integral in terms of area under the graph of a function and as the work done by a variable force. Teachers universally used the first interpretation while only a single American teacher used the second.

The last questionnaire item dealing with the concept of integration asked teachers how they dealt with the sequence of introducing the definite integral and

the notion of the anti-derivative. The vast majority, 84 percent of the Canadian teachers and 91 percent of the American teachers, indicated that they introduced the anti-derivative first. Ten and five percent respectively reversed the sequence. The remaining teachers, about five percent in both cases, did not teach the concept of the anti-derivative.

Major Theorems: The questionnaire probed the presentation of theorems through a set of choices dealing with three major theorems. the teachers' responses are displayed in Figure 4.

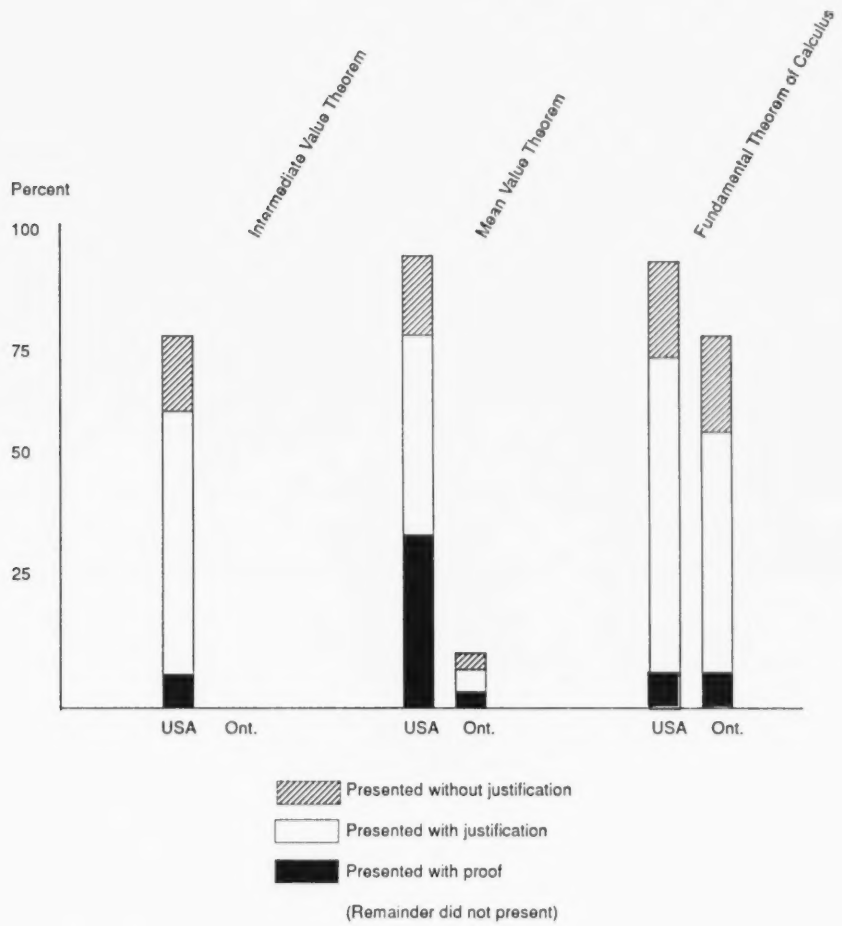


Figure 4. Presentation of three important theorems.

A marked difference between the high school calculus courses in the United States and Ontario occurs for these three theorems. While these three theorems were universally part of the American curricula, this was not the case in Ontario. None of the Canadian teachers included the Intermediate Value Theorem in their courses, and only ten per cent included the Mean Value Theorem. Sixteen percent of Canadian teachers did not present the Fundamental Theorem of Calculus to their students. It is most surprising that this theorem would ever be omitted in a calculus course which, as these did, includes both differentiation and integration

The Learning of Calculus

The learning of calculus at the secondary school level in North America is examined using achievement results on 25 of the test items that were used in the Second International Mathematics Study (SIMS) for the 62 Canadian and 44 American classes discussed above. Of the 25 items, 13 dealt with differentiation topics while 12 dealt with integration topics. The items will be referred to here using their SIMS code numbers. Based on the teachers' reports, all but two of these items, numbers 73 and 122, can be considered part of the curricula of virtually all 106 classes. These

classes. These items test basic material that does not extend beyond the first semester of a typical college course in calculus.

Evaluating achievement results is a complex and often contentious matter. What might appear to be a satisfactory level of performance to one observer might be judged quite inadequate by another. In the final analysis each reader must render his or her own informed judgement. To assist in the process comparative achievement data is provided for several other jurisdictions which participated in SIMS: viz, England and Wales, Japan, New Zealand, and Sweden. In Japan, New Zealand, and Sweden approximately the same percent of the age cohort were enrolled in a course in which calculus was studied, 11 or 12 percent in each case. For England and Wales and the United States the corresponding figures were six and five percent, respectively. For Ontario Grade 13 the figure was 19 percent. Thus, the English and American groups are the most elite which the Canadian group is the least elite. This, obviously, has some relevance on the level of performance to be expected from each group. Achievement levels for the items and for groups of items for each of the six jurisdictions are shown in Table 2.

In Table 2 the differentiation and integration items have been divided into two groups. The items grouped together as Differentiation 1 and Integration 1 deal with basic techniques while the items grouped together in the other two categories deal with simple applications. Considering the entire test, the performance of the Canadian students is the lowest of all of the jurisdictions surveyed. The Americans performed at a level comparable to students in Sweden and New Zealand, but substantially below students in England and Wales and in Japan.

Table 2
Outcomes for 25 Calculus Test Items
(Percent Correct)

Item or Group	United States	Canada (Ontario)	Japan	Sweden	England & Wales	New Zealand
14	82	79	77	72	77	85
72	62	63	56	52	70	58
106	74	63	73	57	80	84
118	61	41	62	34	63	43
Differentiation 1	70	62	67	54	73	68
28	57	56	62	25	50	43
57	24	22	54	16	39	12
88	54	60	54	44	68	65
104	55	46	74	59	59	50
111	29	30	56	43	41	29
112	47	35	62	47	49	32
117	54	58	86	59	80	58
119	49	42	60	39	48	48
122	25	20	56	27	31	37
Differentiation 2	44	41	63	40	52	42
15	85	74	83	71	88	83
73	19	21	51	28	47	25
103	71	58	80	78	81	67
107	80	75	74	53	86	74
109	20	24	67	56	35	14
113	69	56	73	75	78	60
114	39	28	51	25	38	32
116	39	33	41	28	46	33
Integration 1	52	46	56	52	62	49
29	60	54	81	74	67	70
44	27	25	58	59	46	47
58	32	38	66	43	41	26
115	41	26	55	48	40	44
Integration 2	40	36	65	56	49	47
All Items	50	45	63	48	58	49

Achievement in Advanced Placement classes

Due to the rather lackluster achievement of the 44 American calculus classes on the calculus items, it is of interest to know if the 26 classes that were preparing to take the College Board Advanced Placement (AP) exam did any better than the other 18 classes. Williams (1986) addressed this question by analyzing the results on 46 of the test items. His subtest included some pre-calculus items as well as some calculus items that were not included in the above analysis. Since, in the United States, both a pretest and posttest were given, it was possible to determine if there was a difference between the two groups at the beginning of the calculus course. Williams' analysis showed that no statistically significant difference existed in pretest scores. By the end of year, however, his analysis showed that the AP classes had learned more. They scored an average of 14, 8, 5 and 7 percent higher than the non-AP classes on limits and continuity, differentiation, and applications of differentiation subtests, and the total analysis test, respectively. These results were significant at the 0.05 level indicating a high level of probability that AP classes outperformed non-AP classes in the United States on these topics. The AP classes also did better, but not at a statistically significant level, on basic integration. The AP classes did not do better than the non-AP classes on applications of integration.

Williams could not explain, on the basis of his data, why the AP classes outperformed the non-AP classes. This result is, however, consistent with other research that has shown that AP high school students in the United States tend to achieve as much or more than American university students studying calculus (Haag, 1977 and Dickey, 1982).

Summary and Conclusions

There seem to be major differences in the calculus courses as offered in Ontario and in the United States. Many of these differences probably stem from the differences in basic texts being used. In the United States college level calculus text predominate while in Ontario the texts used most frequently have apparently been written specifically for the secondary level.

The American classroom curriculum is generally more extensive at both the beginning and the end of the course. Much more emphasis is put on the foundation areas of limits and continuity, with epsilon-delta definitions and proofs playing a key role in classroom presentations as well as in student learning objectives in the United States. The Intermediate Value and Mean Value Theorems are studied in the United States but not in Ontario. Finally, such topics as arc length, integration by parts, indeterminate forms, and volume of surfaces or revolution are taught more often in American high school calculus courses than in such courses in Ontario.

American teachers who took part in this study emphasized the foundation areas more often and usually expected their students to be able to do epsilon-delta proofs. They did not, however, justify the formulas which they presented later in their course as often as the Canadians did. Canadian teachers were much more apt to give their students a formal proof of the chain rule than the American teachers, for example. There was also a closer relationship in the Ontario classrooms between the level of rigor in teacher presentations and the level expected of students for these formulas.

The use of graphs was employed by the majority of teachers to supplement algebraic presentations. However, while one would expect all teachers to use graphs in discussing the limit of a function at a point and in presenting epsilon-delta arguments, a large number of teachers in both jurisdictions chose not to do so.

The calculator was the one instructional aid, in addition to the basic text, used by the vast majority of all teachers. Only a relatively small number of teachers used movies, models, or computers. Applications of calculus were almost always drawn from the context of the physical sciences and were taken from textbooks or created by the teacher.

The achievement of calculus students in both the United States and Ontario is probably less than satisfactory overall and certainly so for those items dealing with applications of basic concepts. American students tended to outperform their Canadian counterparts scoring five percent higher on the 25-item test. All seven of the items in which one jurisdiction outscored the other by over ten percentage points were in favor of the United States.

A key element in assessing these achievement results is the percent of the age cohort served by these high school calculus classes within each jurisdiction. With this in mind, the overall results achieved in Ontario's Grade 13 classes tend to look better while those in the United States tend to look worse. In Ontario about 19 percent of the age cohort are enrolled in Grade 13 calculus. This is comparatively a very high percentage and can be used to justify somewhat lower achievement results than might be otherwise considered satisfactory. Achievement on the integration items in Ontario still must be considered poor, however.

The American achievement results must cause a good deal of concern when the very elite nature of the high school calculus population is considered as well as the high degree of appropriateness of the 25 items to the

basic course content as reported by the teachers. American results were only slightly above those obtained in Sweden and New Zealand where the percent of the age cohort enrolled in calculus is twice as large as in the United States. The results were considerably below those obtained in England and Wales, and in Japan. Only in the former case is the population about as small a percent of the age cohort as in the United States. Certainly, these results should cause American mathematics educators to reflect on the expectations which exist for calculus instruction in American schools as well as the adequacy of the precalculus instruction which students are currently receiving. The situation appears less problematic where Advanced Placement programs are in place, however.

References

Alexander, D.W. (1987). Personal communication, January 6, 1987.

Dickey, E.M., Jr. (1982). A Study Comparing Advanced Placement and First-Year College Calculus Students on a Calculus Achievement Test. (Doctoral Dissertation, University of South Carolina).

Haag, C. (1977). *Comparing the Performance of College Students and Advanced Placement Candidates on AP Examinations*. New York: College Board.

Henry, P., Jones, C., & Kenelly, J. (1985). The Advanced Placement Program in Calculus. In Hirsch, C., & Zweng, M. (Eds.), *The Secondary School Mathematics Curriculum* (1985 Yearbook of the National Council of Teachers of Mathematics, pp. 166-167). Reston, VA: NCTM.

Wirszup, I. (1980). as quoted in "U. of C. Prof warns of Soviet Science Push." *Chicago Sun Times*, Friday, April 18, 1980.

Williams, J.B., (1986). The Effect of the Advanced Placement Exam on Achievement in High School Calculus Classes. Unpublished paper, University of Illinois.

PARTICIPATION AND OPPORTUNITY TO LEARN AS FUNCTIONS OF STRUCTURAL & ORGANIZATIONAL FACTORS OF SCHOOL SYSTEMS

Edward A. (Skip) Kifer

When international comparisons are made using data from the International Association for the Evaluation of Educational Achievement (IEA) studies, the focus most often is on results of the achievement tests. There is a general interest in knowing which systems do best and which not so well when comparing test scores which, presumably, reflect more or less knowledge and skills in a particular content area.

There are, obviously, other ways to compare these systems. One such way, and the theme of this paper, is to look at how policies and practices of different educational systems distribute opportunities to their students. The results of the Second IEA Mathematics Study (SIMS) are particularly appropriate in this regard since mathematics is perceived as such a crucial subject area in virtually all systems.

What types of students are given what types of opportunities within these varied educational settings is the focus of this paper. Questions of when and how students are selected into different curricula are considered paramount, since that selection determines the kind and amount of mathematics to which a student will be exposed to.

The Samples

Two groups of students, one of students 13 years old and a second that included students in the final year of secondary school enrolled in advanced, university-preparatory mathematics courses, were targeted for the study. Those samples were justified on the following basis: Population A (students 13 years old and typically in grade eight) was chosen because it may be and often is the last time that all students are taking mathematics. Hence, grade eight represents the point where the minimum amount of mathematics is given by a system to all its students. The second group, Population B, is a sample of those students who theoretically have received the most mathematics that a system delivers prior to university or tertiary education. In the SIMS study, these students are the mathematics specialists in the secondary schools of these educational systems.

This paper examines results from both populations. Who participates in what kinds of mathematics available at grade eight serves as an indicator of how opportunities are distributed within an educational system when each student is taking mathematics. How much mathematics is given to how many students and to what kinds in Population B are of interest because they reflect the importance of mathematics to that system.

Results discussed in this paper come from a subset of systems that participated in the survey. The reason for not using results of each system is that the surveys were implemented differently from system to system. Crucial to a discussion of Population A results are measures taken both early in the school year (a pretest) and at the end (posttest). Only eight systems implemented studies with those features. For the Population B section of the paper, two systems (Hungary and the United States) are featured. These two were chosen because they approach differently both the retention of students in school and exposure of students to advanced mathematical knowledge.

The Symbolic Importance of Mathematics.

IEA's second mathematics study was first a study of mathematics — issues of curriculum, students' achievement, and pedagogy were emphasized — but because of the place of mathematics in schools, it could not be only that. With increasing demands for technical expertise coming from the broader social arena, greater weight is placed on mathematical skills and talents than on outcomes from exposure to other traditional content areas. In order to be in greater demand in the job market, or to qualify for more prestigious higher education, a student must navigate the best mathematics a system has to offer. Since the school has a virtual monopoly on such training, the demands on students and schools are obvious. If success in mathematics is a key to later success, and schools determine who gets what mathematics, then mathematics becomes a symbol of modern schooling.

Variation in Tracking Policies - What Mathematics for Which Students?

Because there was, in eight of the systems, a pretest administered at the beginning of the school year, it is possible to describe the allocation of students to classrooms and schools. Average scores between classrooms within schools and between schools reflect policies that are adopted in terms of whether or not students are sorted and tracked into more or less homogeneous classrooms or schools.

If, for instance, students were assigned randomly to classrooms (or systematically assigned to classrooms to insure heterogeneity) within a school, one would expect the average pretest score for each class of students to be about equal. If students were assigned to classrooms according to achievement levels in previous grades or on the basis of an aptitude test, and if the higher scorers were assigned to one class and lower

scorers to another, one would expect average scores to vary greatly between classrooms in the same school. Similarly, if students attended schools on a random basis (or were assigned to schools to promote equity), school averages would be about equal. If, however, there were selection according to prior achievement or if students whose background characteristics correlated with achievement were clustered in separate schools, then one would expect substantial differences between school means.

to both systematic allocation of students to classrooms (despite a provincial policy to the contrary) and different demographic characteristics of the schools. The patterns in France and Ontario show minor differences both between classrooms and between schools, and in neither case are they of the magnitude of classroom differences in the United States or school differences in Belgium Flemish. It appears, therefore, that neither the French nor the Ontario schools have yet begun to sort according to measures of achievement or aptitude.

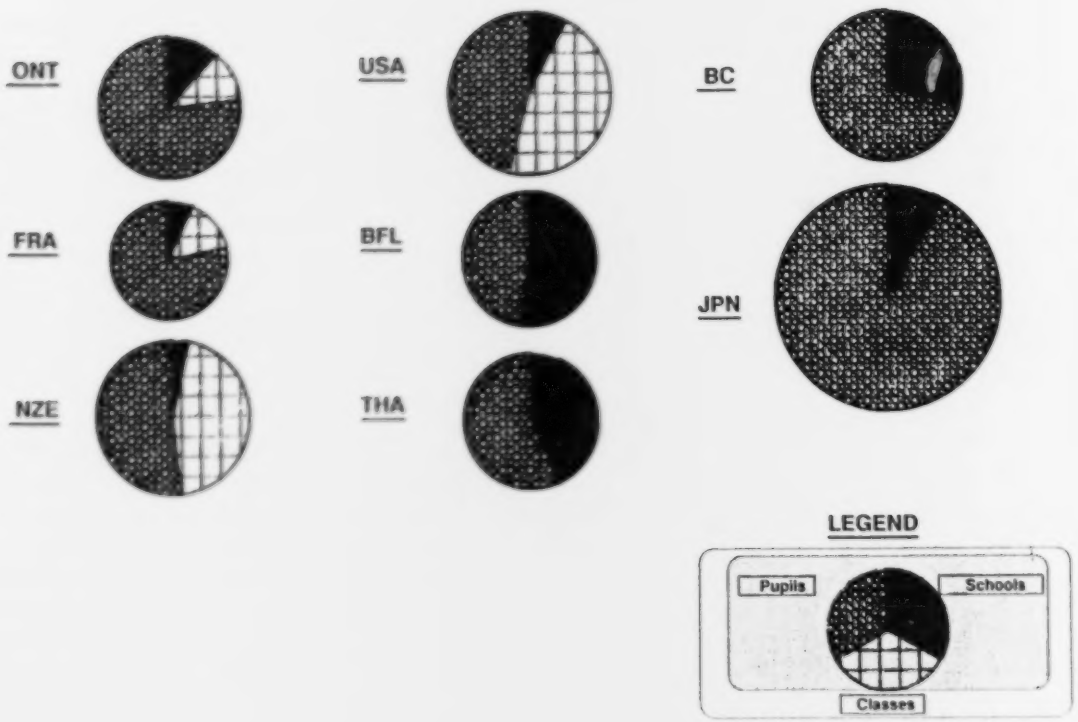


Figure 1. Variance decomposition of Population A Pretest.

Figure 1 contains the results of a variance decomposition of the Core pretest in each of the eight "longitudinal" systems. In that figure, areas of the circles are roughly proportional to the total variance in achievement for each system. The wedges within the circles represent percentages of total variation that is found between students, classrooms, and schools. Those circles which contain only two wedges depict systems that did not sample two classrooms per school. In those cases, the variation is labelled student and school variation although theoretically the wedge for school contains both the classroom and school variance. That is, the between classroom variation, if any exists, is a part of the between school variation not the between student.

Issues of exposure to instruction and participation in mathematics in Population A are tied to policies of how students are allocated to classrooms or school. If there is a common curriculum and no attempt is made to place particular students in special classes, then participation and exposure are more or less common for each student. If there is some kind of tracking within the system, then questions can be asked about whether decisions to track are related to the kinds of mathematics experience students are given.

The differences between systems are dramatic. Not only are the total variances (individual differences within a system) of strikingly different magnitudes, but also that variation is divided (how individual differences are responded to) in distinct ways. In Japan, for instance, almost all of the large total variation is between students. Since there is such a small amount of between school variation, variation between classes in the same school must likewise be small.

One can infer that the Japanese either ignore individual differences when assigning students to classrooms, or they implement policies that produce equality among classrooms and schools. There is no homogeneous grouping in mathematics in Japanese schools at this grade level and there appears to be no sorting by school.

In bold contrast to the Japanese pattern of variation stands that of the United States. The magnitude of the between classroom component in the latter is its single largest component and exceeds comparable values in all of the other systems.

Other systems, too, have distinctive patterns. New Zealand, despite the fact that it purports to have a national curriculum, reflects a pattern very similar to the United States. The between school differences in Belgium Flemish are a reflection, one assumes, of the fact that there are different school types (vocational, general and technical) and different organizing authorities for students at this grade level. The between school differences in Thailand can be hypothesized to reveal differences between urban and rural schools, while those in British Columbia apparently are related

The Belgium Flemish, New Zealand and United States of America Cases.

It is obvious from Figure 1 that different systems have different policies insofar as the allocation of students to classrooms or schools is concerned. In another paper (Kifer, in press), I have done detailed analyses of the nature and consequences of such policies in Belgium Flemish, New Zealand and the United States. Here I will highlight those findings rather than portray them in detail.

Different Types of Tracking Have Different Consequences

Belgium Flemish at this grade level has different types of schools for its pupils, and those pupils are exposed to different amounts of mathematics. The United States has different types of mathematics classrooms within each school, and in those classrooms students are exposed to radically different types of mathematics. In New Zealand schools, students are sorted into classrooms by, apparently, measures of previous achievement and then given either more or less mathematical content.

The most dramatic example of how tracking policies influence what mathematical content students are exposed to comes from the United States. Figure 2 is a series of Box and Whisker Plots which describe, by four distinct classroom types in the United States, teachers' ratings of the Opportunity to Learn (OTL) the mathematics reflected by the SIMS achievement test. OTL was gathered by asking each teacher to look at each test item and decide whether or not the material needed to answer the question correctly had been taught to the students.

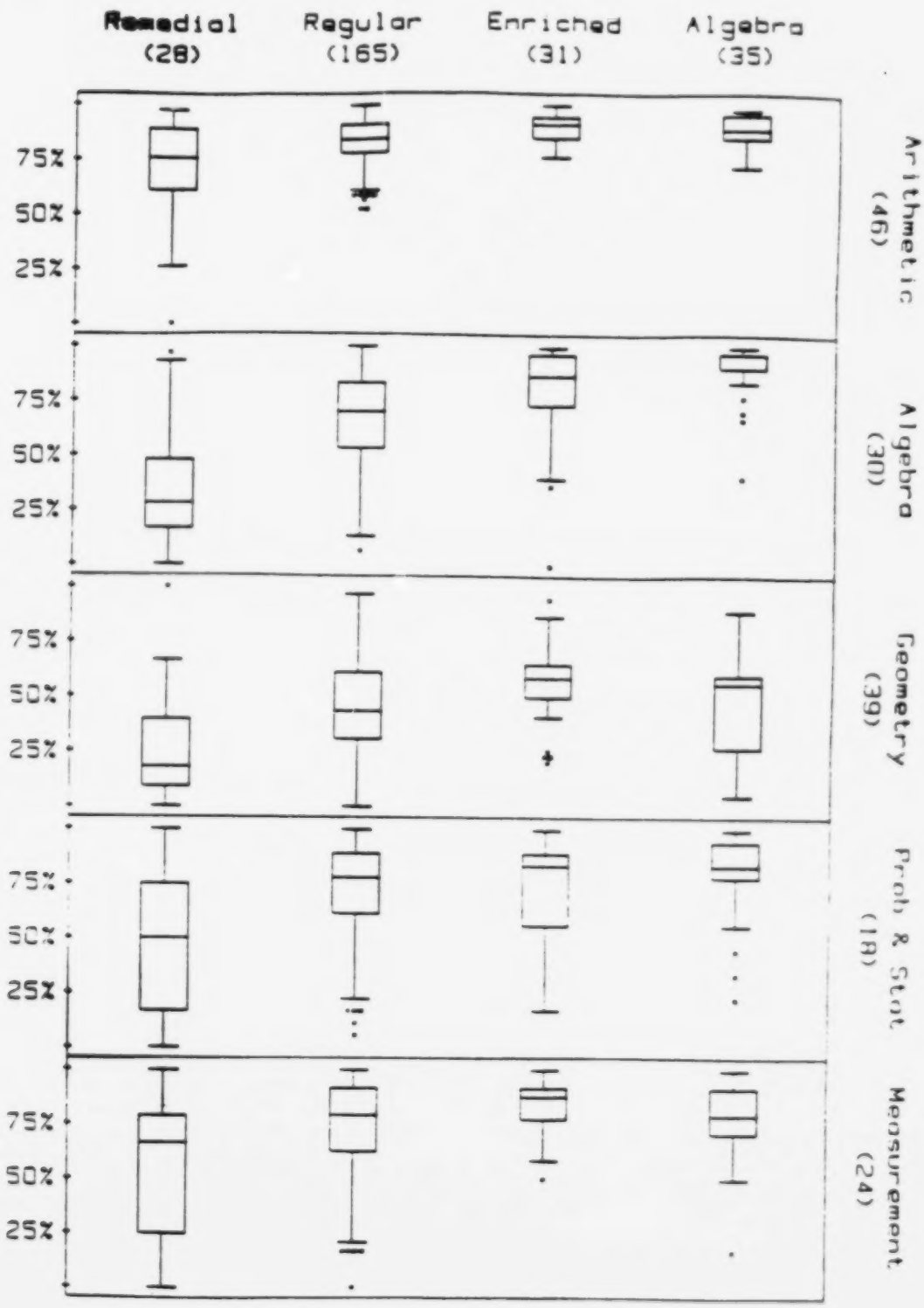


Figure 2. OTL by content area for United States of America class types.

It is evident from Figure 2 that sorting students and differentiating the curriculum are two sides of the same coin. Those students, for example, who are in Algebra classes (the high track in the United States) are exposed to very different kinds of material than those in other types of classrooms.

Though not nearly so dramatic as what is found in the United States, tracking of students leads to different types of exposure to mathematics in both Belgium Flemish and New Zealand as well. Those differences, however, are both smaller in magnitude and of a different kind. In those two systems, students in "better" tracks tend to be exposed to more mathematics.

The Sorting Is Inefficient

In each of these three systems it can be assumed that procedures used to allocate students to classrooms are meant to be rational and efficient. The analyses suggests, however, that if these systems are operating meritocratically — that is, it is desired that the best students be in the highest tracks and vice versa — they are not doing very well.

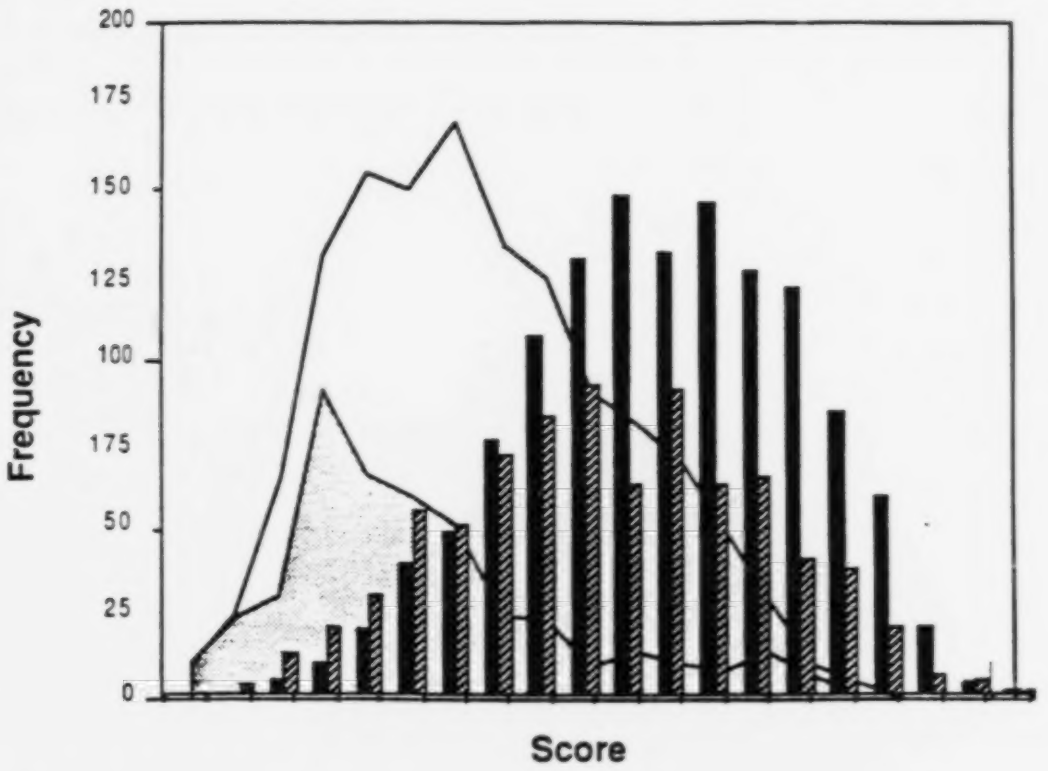


Figure 3. Distribution of pretest scores by school type in Belgium Flemish.

Figure 3 shows the distributions of pretest scores by school type in Belgium Flemish. What is worthy of note is that a substantial number of students in vocational and technical schools have pretest scores on the SIMS test that are above the average for the traditional and comprehensive schools. Hence, if the selection were done by the system (as opposed to individual choice) and based on merit, quite a number of students have been mis-classified.

In the United States, of those students in the top 10 percent of the distribution of pretest arithmetic scores, only one-half are placed in algebra classrooms. Of the students in the top quarter, slightly less than one-third were in algebra.

For New Zealand, students who appeared to be high scorers in one school would be placed among the low scorers in another. Hence, they too were making a substantial number of classification errors if merit or prior performance were the means for students to get preferable curricular experiences.

The Tracking Has Social Consequences

Not only is tracking inefficient and error prone but the practice also has social consequences. Analyses (Kifer, 1984) of whether there were background characteristics of students related to participation in the various tracks indicated social biases in that allocation.

Figure 4 depicts the relationship of Father's and Mother's educational levels and whether a student was in a high or low scoring classroom in New Zealand. The high scoring classrooms had a disproportionate number of students whose fathers or mothers were highly educated. Conversely, low scoring classrooms were disproportionately populated with students whose fathers and mothers had lower levels of education. In the United States, students who are white, female, and come from wealthy homes are placed in the favored tracks. Those who are not white, are boys, and are not wealthy are more likely to be placed, regardless of test score, in the lower level classes. Class and gender effects are present in Belgium Flemish but to a much lesser degree than what is found in either the United States or New Zealand.

The Cases of France, Japan and Ontario.

It is not the case that some systems track or sort students and others do not. It is a matter of when the sorting occurs not if it will occur. Yet, the systems of France, Japan, and Ontario have in place, apparently, policies which attempt to insure that virtually all students are exposed to common material at the Population A level.

Remembering that this population was chosen because in most systems it is the grade level where all students still take mathematics, these three systems have chosen to make the educational experiences of the young common ones in mathematics. Later, each will sort.

This egalitarian approach to mathematics in France is a result of national changes instituted in the educational system during the late 1960's. Concerns were expressed at that time about the lack of common opportunities available to students of this age cohort. Selection into types of curricula occur in France during the upper secondary school, rather than during this relatively early period of a student's school life. These results suggest that the new system gives more students a more equal chance of going in the most desirable educational route by guaranteeing equal opportunities through the elementary school years.

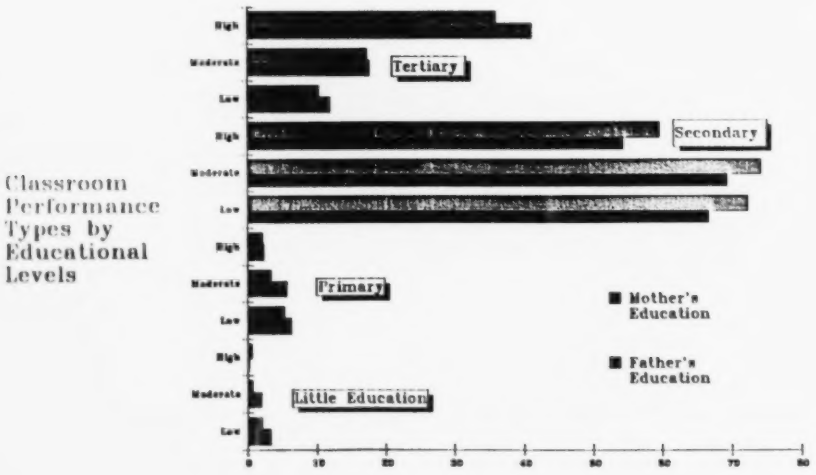


Figure 4. Percent of students by classroom type and educational level of parents in New Zealand.

For Japan, whose sample is one grade level earlier than others in this set, entrance into upper secondary school is the demarcation of the change from common opportunities to differentiated ones. These decisions — which students enter which type of schools — occur about three years later than this grade level and are based primarily on entrance examinations.

For the Ontario system, sorting occurs at the next grade level. As students enter the secondary school, a number of different types of measures are used to determine which curricula they will participate in. The extent to in which background characteristics of students are related to participation in the most favored curricula obviously cannot be addressed with these data. A data set for the subsequent year would be needed to address these problems.

Summary

Population A systems provide a contrast between those which have more or less egalitarian policies (France, Japan, Ontario) versus those with meritocratic one (Belgium Flemish, New Zealand, USA). Which is the preferred set? Some would argue that the merits of egalitarian versus meritocratic educational practices should be found in differences in achievement rather than differences in opportunities or in equality of participation. Previous IEA studies suggest that comprehensive schools do not negatively affect the performance of the most talented. And, selective schools do not necessarily enhance the performance of those who are enrolled there. Such analyses, however, have been based on older populations of students and may or may not be appropriate in this context.

There is little, if any, direct evidence of the efficacy in terms of achievement of either the egalitarian or meritocratic approaches and practices among the participating systems. Since these are national systems and this is a sample survey, variables which may operate to produce high or low performance and which distinguish between the systems or the contexts in which they operate, are simply not available. It would, therefore, take an extremely strong inference to state that in terms of cognitive achievement, as measured by the SIMS test, there is a decided advantage of one set of practices over the others.

Nevertheless, there may be findings and indirect evidence within the study that would allow one to prefer the practices of the egalitarian systems — Canada Ontario, France and Japan — over the others. First, Population A students in both France and Japan scored well on the cognitive tests and showed rather remarkable gains on subsets of the items. And, in previous analyses, it was shown that Canada Ontario, which is both comparable in terms of variance and achievement to the United States, shows slightly greater growth than does the United States. In addition, the patterns of gain for the two systems are very similar. Hence, straightforward comparisons, though arguably weak by nature of the design of the survey, show superiority on the part of egalitarian practices.

Logic, too, supports these egalitarian policies and practices. If a system wishes to select the most talented students and provide them with the best educational opportunities, then the longer that the selection is put off, the better it will be.

The sorting of United States students, for instance, starts much earlier than the Population A grade level where the tracks are firmly in place. If a mistake in selection is made prior to grade eight, the child's school career is obviously affected. And, there are no systematic ways, even if the child has the required talent, to rectify the mistake. The child could be very good but still be in a low track because early in his or her career an error was made. If, however, there were no tracking or selection in the United States, and there were no concomitant differentiation of curriculum, no opportunities would have been thrown away. Hence, the longer a system waits to sort the more likely it is to have a developed (in the talent sense) an identifiable cohort on which to sort. Since these three systems — Canada, Ontario, France and Japan — have not yet sorted, there practices are preferred to those of other systems because they up until now have made fewer errors in the selection process.

Population B

As will be shown later policies adopted at the Population A level influence profoundly what can be done at Population B. Yet, the issues of participation and exposure to mathematics content are different for the two populations.

Virtually all students are taking some type of mathematics in the 13 year old population; by the end of secondary school, depending on the system, either a large proportion of the cohort is no longer enrolled in school or not taking mathematics or both. Figure 5 shows those proportions. The estimated percent of the cohort still in school ranges from a high value of over 90 in Japan to a low of 17 in New Zealand and England. The percent of the cohort taking advanced mathematics courses ranges from a high of 50 percent in Hungary to lows of 6 percent in Israel and New Zealand.

Country	In School	In Advanced Mathematics
United States	82	13
Sweden	24	12
Scotland	43	18
Ontario	33	19
New Zealand	17	11
Japan	92	12
Israel	60	06
Hungary	50	50
Finland	59	15
England	17	06
British Columbia	82	30
Belgium Flemish	65	9.5

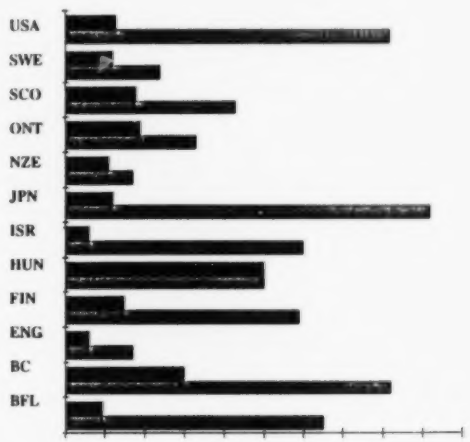


Figure 5. Participation in schooling and advanced mathematics: Population B.

The United States has a relatively high rate of retention (82 percent which is second to Japan) and is in the middle in terms of the proportion of the cohort taking advanced mathematics.

Across these systems, two phenomena are evident. First, there has been a selection made across the student cohort. That is, not all students progress through these systems until they reach the terminal year of secondary schooling. Depending on the system, student attrition can be a matter of dropping out of school and entering the job market or it can be that there is another, earlier, school leaving point where the majority of students get a certificate and leave school having completed the required number of years. In the latter systems, a minority of the cohort continues secondary school in order to prepare for university. Second, among the students who remain in school, it is a fraction of them, in most systems, who are taking the most advanced mathematics offered by the secondary school. Also, among these students there is a possibility that they are taking mathematics, but not at the highest level. These educational systems vary dramatically in the policies that determine which students remain in school and, of those, which continue to take advanced mathematics. The section below focuses on two extreme cases of dealing with these issues.

The Hungarian Example

While most systems are very selective at this level, a striking exception is the Hungarian case. While having "only" 50 percent of the cohort still in school, all of those are taking advanced mathematics. This finding suggests that very different policies inform the mathematics community in this system. One conjecture would be that the Hungarians do not believe they can afford to have mathematics be an elitist content area. Mathematical knowledge is sufficiently important to be a part of each student's experience at this level.

Miller and Linn (1985) examined achievement patterns in light of the different retention rates in these systems. They report two things that are relevant to the Hungarian system and this paper. The first is that the average level of achievement for Hungary's students is close to the bottom among the systems; the second is that the top 1 percent and top 5 percent of Hungarian students perform near the top of the distribution of scores for these systems. From an international perspective it appears that the Hungarian experience allows them to have it "both ways." Not only are they providing advanced mathematical experiences to a large percentage of the cohort, and thereby increasing dramatically the sum of mathematical knowledge in the culture, but also they are doing it without sacrificing the talents of their most capable students. As a model for providing both opportunity and creating a pool of talent, Hungary's bears scrutiny.

The Case of theUnited States

The situation in the United States is practically the opposite of the Hungarian one. In the United States there is a high retention rate but a modest percentage of students taking advanced mathematics.

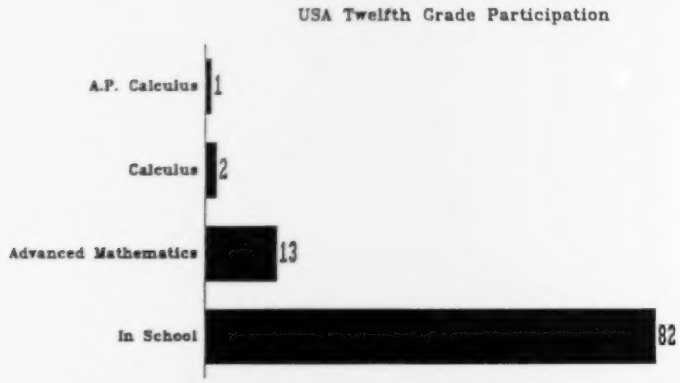


Figure 6. Percent of the United States cohort in school and in mathematics.

The latter count, however, is misleading. There are, in fact, differentiated curricula at this level as well. Figure 6 shows how the United States stands when content areas are broken broadly into calculus and other courses. The results suggest that there is a rather smaller cohort in the United States than in other systems. Since calculus is standard fare for these other systems, the United States percentage is really much lower than it appears. The calculus courses are further differentiated between those that are considered Advanced Placement and others. The numbers of students who are enrolled in Advanced Placement Calculus is extraordinarily small; it is estimated that it is less than one percent of the cohort. So the American elite is very small and a far smaller proportion of students in the United States is receiving mathematics experience comparable to that of students in other systems.

Conclusions

Systems which track students early profoundly affect the chances of many students being exposed to much of the mathematics that is offered to students in other educational systems that put off tracking until later. By grade eight in the United States, for example, less than 15 percent of the students are in a track that will allow them to take calculus in Grade 12. By grade twelve another 10 percent or more (of the cohort) has dropped out so that there is little participation in advanced mathematics in the United States compared to that in other countries.

Not only is participation in advanced mathematics low in the United States, but combined with the Population A findings, there is a serious question of whether the most talented students are enrolled in the advanced mathematics courses. If one half of the top ten percent of students are taking courses in grade eight which allow them to take the most advanced mathematics in grade twelve, it is conceivable (though not proven) that the students who do take the most mathematics are not the best ones. The best ones may have been selected out by errors of early tracking.

The Hungarian system shows another approach to educating students mathematically. Although its retention rate is lower than most other systems (50 percent of the cohort still in school), since it does not differentiate its mathematics curriculum, it has much higher participation than other systems. Apparently in Hungary mathematics is considered important enough to be offered to a large percentage of the cohort.

Selection Effects

The fact that early tracking differentially affects the genders, persons from different social classes, and different ethnic groups raises additional issues. Two not so easily answered questions are raised by these differential participation rates. The first has to do with the issue of equity in general. Talented students who are poor and from minority backgrounds are being excluded from fullest participation in school mathematics. This loss of human resources has implications for the knowledge of mathematics that informs a culture, but also raises moral issues.

The second issue is what to do about the first. SIMS provides results that identify the problem but, as is the case for many such projects, does not provide a basis for solving it. Because it is an international survey, and because these systems are quite varied in terms of their policies, there are different models available to those who wish to change how students are educated mathematically.

The Problem is Participation

It is interesting to note that by Grade 8 in the United States enough sorting of students has occurred so that the percentage of students taking algebra is about equal to the percentages that take the most advanced mathematics offered at Grade 12 by educational systems in other countries. The tracking there is so rigorous that, in fact, it is assured that participation in advanced mathematics in going to be small in secondary schools. But these other systems are selec-

tive as well. To have but 10-15 percent of a cohort experiencing the best a school system has to offer in mathematics is by no means exceptional. Is not good mathematics more important than to be offered to such a limited number of students? It appears to this writer, that participation in the best a school has to offer is a major issue for each of every one of these systems.

References

Kifer, E.A. (1984) Issues and Implications of Differentiated Curriculum in the Eighth Grade. National Conference on the Teaching and Learning of Mathematics in the United States. Atherton Conference, Monticello, Illinois.

Kifer, E.A. (in press) Opportunities, Talents and Participation L. Burstein, (Ed.) *Second International Mathematics Study: Student Growth and Classroom Process in Lower Secondary Schools*.

McKnight, C. C. et. al., (1987) *The Underachieving Curriculum: Assessing U.S. School Mathematics from and International Perspective*. Champaign, Illinois: Stipes Publishing Company.

Miller, M.D. and Linn, R.L. (1985) *Cross-national achievement with differential retention rates*. Urbana, Illinois: mimeo.

CONTENT REPRESENTATION IN MATHEMATICS INSTRUCTION: A CASE STUDY OF THREE COUNTRIES

Curtis C. McKnight · Thomas J. Cooney

A characteristic feature of mathematics instruction is that its mathematical content can be represented in a variety of forms. These forms often differ widely in their complexity. Further, they differ in the ease with which they may be comprehended and in the connections that may be made to existing cognitive structures of learners.

For instance, when teaching the concept of common fractions, teachers can interpret such fractions, among other ways, as parts of a region compared to the whole of the region (presented as a figure divided into equal parts with some parts shaded and others not); as a division of integers; as related to physical measurements such as length, area or volume; or as a corresponding fraction in decimal form. Certainly these various representations for the fraction concept would have differing references to structures of existing knowledge for various learners. These representations vary in the degree to which they rely on more perceptual, iconic elements or on more abstract, symbolic elements. These representations are thus likely to be processed quite differently by different learners.

An essential element of the pedagogical task in mathematics is, then, the choice of one or more representations for the content to be taught, whether this decision is made by the teacher directly, by a group creating a curriculum guide, or by the authors of a textbook. In any case, the teacher is the final arbiter of the pedagogy used and has the possibility of choosing content representations to supplement or replace those received from other sources.

The Second International Mathematics Study's (SIMS) questionnaires on classroom processes for specific content areas yielded rich, detailed descriptions of the instruction provided for selected topics in the areas surveyed. The descriptive wealth of the data from these questionnaires offers the potential for casting considerable empirical light on questions about content representation in mathematics instruction.

The authors have taken the approach of examining "local" clusters of related information for selected subtopics (e.g., the concept of common fractions, the addition and subtraction of integers, finding the area of a parallelogram, etc.), rather than a strategy of looking at data at a more "global" level of topics which combine several subtopics (such as arithmetic, algebra, measurement, etc.). Aggregation to more global "topic" levels often involved a confounding of any explanatorily interesting classifications. The results of these investigations appear elsewhere.

There are many approaches to studying content representation strategies as implemented in mathematics instruction. The most obvious would be to study the specific content representations implemented by teachers in various educational systems for various topics and instructional settings. Such an investigation of specifics would be profitable but, used to study a set of more than twenty subtopics available in the SIMS classroom process data, it would involve examination of a complex array of options implemented in an equally complex range of instructional settings set in a context of inter-classroom comparisons within each system investigated and of multi-system comparisons. Specificity in the study of content representations is obtained at the price of large increases in the complexity of the phenomena to be explicated.

It seems reasonable that the likelihood of identifying essential structures and relationships in a set of phenomena is at best inversely proportional to the complexity of those phenomena. If variables that simplified the phenomena without destroying their essential features could be attained, they should increase the likelihood of finding significant structural relationships.

While this generalizing strategy was adopted for the more extensive investigations reported elsewhere, it seemed worthwhile to check the assumption of the value of this approach by seeking an opportunity to analyze at least one small topic area in all its specificity, to examine carefully the descriptive power of such a concrete approach, and to assess more directly through such an example the trade-offs between generality and specificity. The present study is an attempt to do this.

The discussion which follows examines only one subtopic — that of common fractions instruction. It restricts attention to the educational systems of three countries — France, the United States and New Zealand. These systems were chosen because they provided clear contrasts in instructional approach for the mathematical topic chosen. In France, instruction on common fractions is largely delayed to the grade containing students about age 13 (Population A in SIMS), while such content is introduced much earlier in both the United States and New Zealand, but in quite different ways. Selection of this topic restricted use of

SIMS data to that for Population A, classrooms at the grade level at which the median mid-year age was 13. These restrictions have made possible a somewhat detailed and specific look at how instruction in common fractions is carried out by the teachers in these three systems.

Resource and Time Use in Fraction Instruction

Among the first concerns in instruction on any mathematical topic is whether the topic is to be treated as a new or review topic, how much time is to be allocated for instruction on the various aspects of that topic, and which resources are used in providing that instruction. SIMS data is used here to compare France (FRA), New Zealand (NZL) and the United States (USA) on these components of instruction in the concepts of and operations with common fractions.

Teachers were asked whether various aspects of common fractions instruction was taught as new content, reviewed and then extended, reviewed only, neither taught nor reviewed because it was assumed prerequisite knowledge, or not taught even without such an assumption. Figure 1 presents these data.

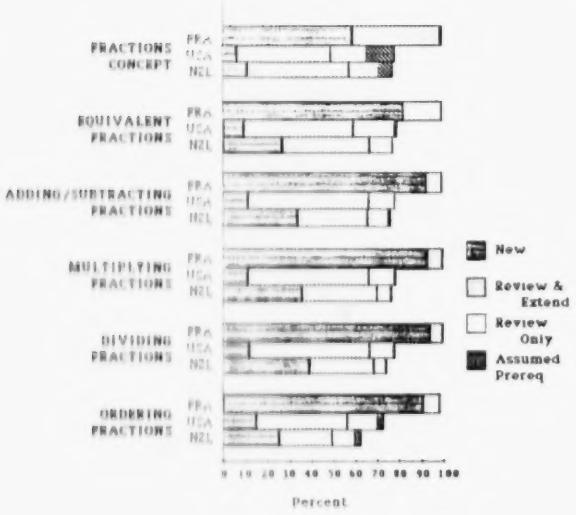


Figure 1. New vs, Review Instruction for Six Subtopics in Three Countries.

Figure 1 shows that for classrooms in France, almost all aspects of this material was presented as new content (which accords with national reports of the mathematics curriculum in France). In New Zealand, this material was often presented as new content but about equally often reviewed and extended. This suggests less uniformity in New Zealand's curriculum in this area or the existence of two or more streams in the curriculum. In the United States, a small proportion of classrooms presented this content as new materials while most reviewed and extended it or reviewed it only. This accords with the fact that there were four types of programs identified at this grade level in American schools and only one of them, remedial classrooms, often treated as new content this material which had been in the curriculum for some years.

Just as the three systems differed in whether this content was treated as new or review material, they also differed in the amounts of time allocated to it. Figure 2 presents "box and whisker" plots of the distribution of time (in minutes) allocated to common fractions subtopics. In such a box and whisker plot, the box runs from the 25th percentile to the 75th percentile, with the line inside the box indicating the median. The lower "whisker" ends at the 5th percentile and the upper "whisker" ends at the 95th percentile. The box thus encloses the middle 50 percent of the distribution and the whiskers enclose the middle 90 percent. Figure 2a presents the total time indicated for common fractions instruction while Figure 2b presents the time for the same six aspects of fraction instruction presented in Figure 1 plus an additional aspect, time devoted to applications and problem solving related to fractions (textbook word problems, problems related to real world situations, etc.).

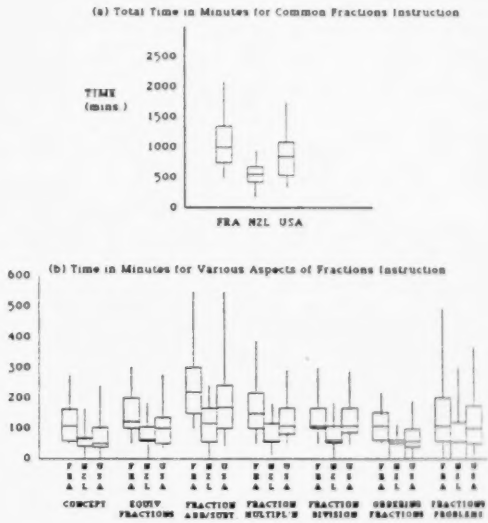


Figure 2. Distribution of Time in Minutes Spent on Common Fraction Instruction.

From Figure 2a it can be seen that the least time was allocated to common fractions instruction in New Zealand and the most in France (where it was essentially a new topic). While there was considerable uniformity among the time allocations in New Zealand, there was considerably more diversity in both the United States and France.

Figure 2b shows the spread of time allocations received by the seven aspects of fraction instruction. It can be seen that the addition and subtraction of fractions received relatively more attention in all three countries. The addition and subtraction of fractions also showed the greatest diversity in time allocations, followed closely by problems and applications of fractions. In all cases, France allocated more time than did the others and New Zealand allocated less. The pattern reflected in the overall time allocations in Figure 2a was consistently reflected across the seven subtopics of Figure 2b.

Teachers in the three systems also differed in the resources used for fractions instruction. The SIMS instruments distinguished between primary resources (those used frequently) and secondary resources (those used occasionally). Data were gathered on six categories of resources, any of which might be used by an individual teacher as either a primary or secondary resource.

The primary resource used by most teachers in all three countries was the student textbook. Other published textbooks and materials (workbooks, worksheets, etc.) were an important secondary resource in all three countries, although they served as a primary resource in only 10 to 20 percent of the classrooms. American teachers made slightly more use of both kinds of text materials than did teachers in France and New Zealand. Locally produced text materials were also an important secondary resource, and in France they were a primary resource for almost half the classes (significantly more than in the other two countries). By comparison, the other categories of resources (commercially or locally produced individualized materials; commercially or locally produced films, filmstrips, or teacher demonstration models; and commercially or locally produced laboratory materials for student use) were little used. While they served as a secondary resource for small percentages of teachers in the United States and New Zealand, they were virtually unused in France. New Zealand made somewhat more use of laboratory materials as a secondary resource than did the others.

A Look at Content Representation

One of the more interesting features of the SIMS instruments which gathered data on classroom processes were questions that examined the use of each of an array of content representations during instruction for specific subtopics. Part of the information gathered was whether a particular representation was emphasized ("used as a primary explanation, referred to extensively or frequently"), used but not emphasized, or not used at all.

For example, one question gathered data on the use of each of ten content representations for instruction on the common fraction concept. These data are presented in Figure 3. It can be seen that the representations most frequently used or emphasized in all three countries were fractions as decimals, fractions as quotients, and fractions as parts of regions. While about half of the teachers in all three countries emphasized fractions as parts of regions, considerably more of the teachers in the United States emphasized fractions as decimals and fractions as quotients than did those from the other two countries.

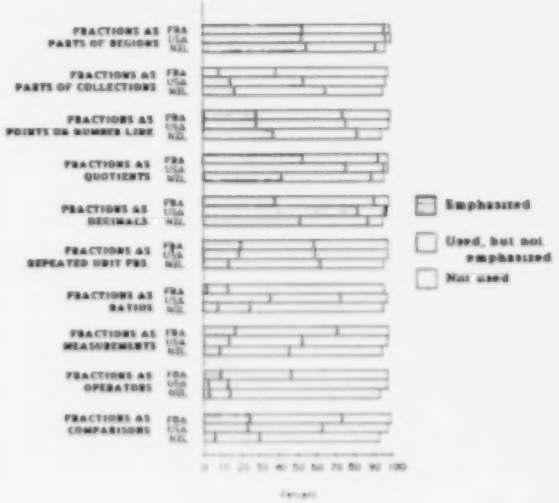


Figure 3. Representations for Common Fractions Emphasized, Used and Not Used in Instruction on the Common Fraction Concept.

Few other representations received emphasis by 25 percent or more of the teachers in a country, although several others were seen to have considerable use but not emphasis. Representing fractions as the coordinates of points on a number line received a fair amount of emphasis in all three countries and especially in New Zealand. Interpreting fractions as ratios was

emphasized by over 30 percent of the teachers in the United States, but very little in the other two countries. Representing fractions as comparisons was emphasized by about 25 percent of teachers in both France and the United States but received considerably less emphasis and use in New Zealand.

In summary, it appears that there are both important commonalties and important differences in the content representation patterns for instruction related to the common fraction concept. A core of three representations were the most often emphasized with a few others supplementing this core for at least some teachers. The range of representations emphasized seems fairly narrow, while the range of representations used but not emphasized was considerably wider.

A second question gathered information on interpretations of the addition of fractions. Interpreting the sum of two fractions as the union of two regions was emphasized by about 30 percent of the teachers in both France and the United States and used by about another 40 percent. Interpreting the sum of two fractions as the sum of two quotients was emphasized by over 30 percent of the French teachers and received considerable use in all three countries. In comparison to interpretations of the common fractions concept, very few of the interpretations of adding fractions received much emphasis in any of the countries and even the use of various interpretations was relatively more restricted. This is at least suggestive that a richer array of content representations are brought into play for more conceptual topics than is the case for more procedural topics.

Additional data was gathered about instruction on the addition of fractions. One question sought to determine which procedures for addition of fractions were emphasized and used in the various educational systems. Six procedures were considered in the instrument — using the least common denominator (LCD) in a horizontal format, using the LCD in a vertical format, using any common denominator in a horizontal format, using any common denominator in a vertical format, using a formula such as

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$

or using transformation to and addition of equivalent decimals.

There were some important differences among the countries. Using the LCD in a horizontal format received extensive emphasis in both France and the United States but relatively less in New Zealand where, instead, using the LCD in a vertical format was emphasized more often (and using a vertical format with any common denominator was used far more often than by either of the other two countries). Using decimals was emphasized frequently in all three countries but somewhat more often in New Zealand. Thus, there were fairly distinctive national patterns in the procedures developed for adding fractions, distinctive patterns that were less characteristic of the content representations chosen.

Data were also gathered on the techniques used by teachers in teaching the addition of fractions. Three possibilities were considered — presenting only numerical examples to demonstrate the procedure, using numerical examples first and then presenting the procedure symbolically ("example then rule"), or presenting the procedure symbolically and then illustrating it with numerical examples ("rule then example"). Patterns characteristic of the three countries stood out quite clearly. Few teachers in any of the countries made much use of the "deductive" approach of presenting the general rule and then presenting numerical examples. About 75 percent of the teachers in the United States presented numerical examples only while over 80 percent of the French teachers used the somewhat more formal approach of presenting numerical examples followed by stating the general rule or pattern. The teachers of New Zealand showed somewhat more diversity, with just over half presenting numerical examples only and a fair proportion presenting numerical examples followed by the general case.

A Second Look at Content Representation

Clearly there are many approaches to studying content representation strategies as implemented in mathematics instruction. While the most obvious approach would be to study the specific content representations implemented, it would involve a bewildering combination of complexities when applied to many cases. Variables that simplified the phenomena without destroying their essential features should increase the likelihood of finding significant structural relationships.

For instance, the examination of the number of content representations used in a given instructional setting, rather than the specific representations used, offered parsimony and the possibilities of greater generalizability and explanatory power, but at some risk of missing relationships tied to the specifics of the situations. Thus, one characteristic of interest was simply the number of content representations used by each teacher in instruction related to a subtopic. This was captured in a variable, VARIETY, a simple count of the number of the different content representations emphasized or used.

A second example of interest was the relative balance in instruction on a subtopic between representations which emphasized in their form more perceptual elements (e.g., shaded regions for interpreting fractions) and those which emphasized more symbolic forms (e.g., fractions as divisions). The relative balance

in instruction on a subtopic between perceptual form representations and symbolic form representations, was indexed by a variable, BALANCE (OF EMPHASIS), which was calculated by taking the proportion of symbolic emphases used (that is, the number of "symbolic" interpretations used, divided by the total number of possible symbolic representations on the list for that subtopic) minus the proportion of perceptual emphases used (that is, the number of more perceptual representations used divided by the total number of possible perceptual representations). BALANCE, defined in this way, took numerical values from –1 through +1. A positive value indicated relatively more emphasis on the symbolic, a negative value relatively more emphasis on the perceptual, and a value close to 0 indicated relatively balanced use of both perceptual and symbolic emphases.

Alternative, more restricted counterparts to VARIETY and BALANCE could be obtained by the same quantifying operations using only those representations which were emphasized and not those which were used (but not emphasized). These alternative definitions might give a very different picture of the "heart" of content representation than that provided by the more inclusive definitions.

The data showed that a relatively large number of representations (5-8) were used in all three countries. The United States showed somewhat greater diversity of use. In comparison, all three countries emphasized a far more restrictive set of representations, with France showing slightly greater diversity in representations emphasized.

A sense of these data can be given by graphing the percent of teachers in each country who use the various numbers of representations possible (0 to 10 for instruction on the common fractions concept).

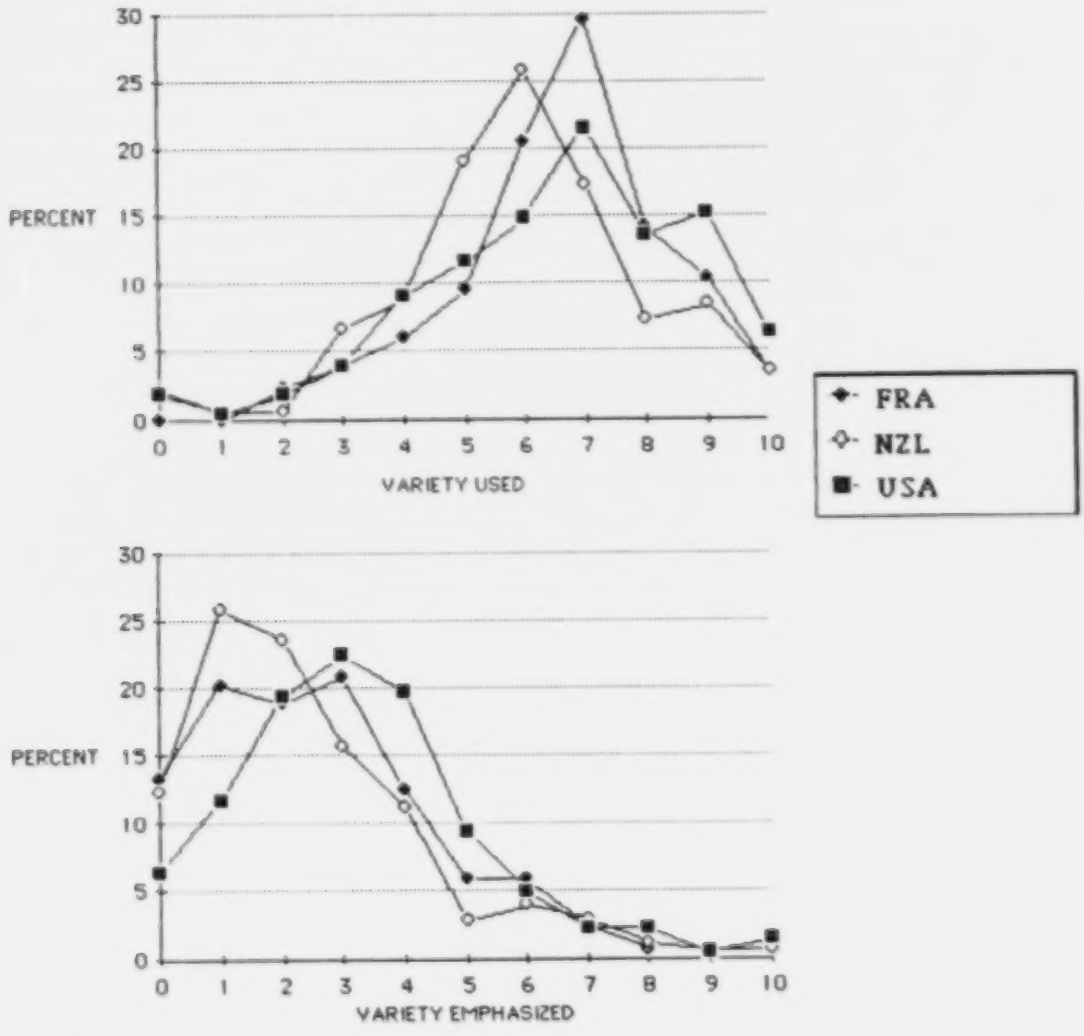


Figure 4. Variety of Representations Emphasized or Used in Three Countries for Common Fractions Concept Instruction.

Figure 4 presents such graphs, both for the VARIETY of representations used and emphasized. From Figure 4 it is clear how much difference there is between the VARIETY emphasized and that used, and how New Zealand differs from the other two countries.

A similar analysis showed both the similarities and differences in the VARIETY used and emphasized for instruction on common fraction addition. The VARIETY of representations used is somewhat more "diffuse", i.e., more spread out and less "peaked" for instruction on fraction addition in comparison to the fraction concept instruction. However, the VARIETY of representations emphasized was very restricted for fraction addition. For New Zealand and the United States, the modal number was zero, i.e., no representations were emphasized by over half of the teachers in those countries. For France the modal number emphasized was one. Thus, there was a marked and suggestive difference between instruction for the conceptual and procedural aspects of this topic.

Figure 4 showed some of the benefits of abstraction in comparison to the more specific data on representations presented in Figure 3. Another way to examine this trade-off between specificity and abstraction more directly is to create something like a "power curve" for each representation. This is done by plotting the percent of teachers in each country using or emphasizing that specific representation for each level of the VARIETY variable. Figures 5 and 6 present such graphs for the common fraction concept for two of the countries.

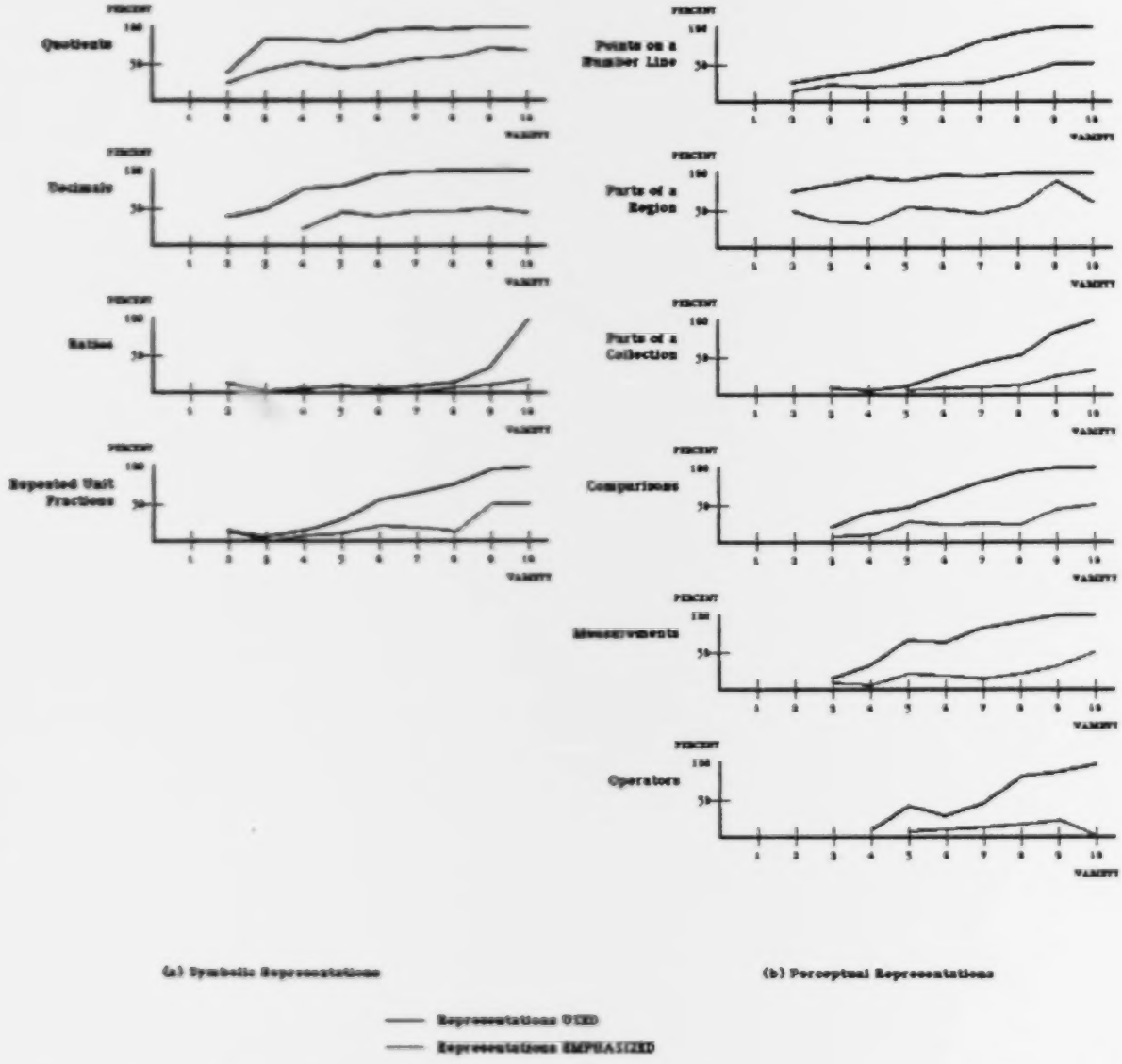


Figure 5. Percent Using and Emphasizing Specific Representations for Common Fractions Concept by Differing Variety for France.

The ten representations available for the common fractions concept are categorized into two groups — symbolic and perceptual representations. The left column of each figure contains curves for the four relatively more symbolic representations and the right column those for the six relatively more perceptual representations. Each graph contains two curves — an upper, black curve for VARIETY of representations emphasized or used and a lower, gray curve for VARIETY of representations emphasized. By the nature of the case, all teachers with a VARIETY used or emphasized score of 10 have used or emphasized all listed representations and thus each of the upper, black curves must end at the maximum of 100 percent. Such is not the case for the lower, gray curve. The height of each curve and how "early" (how far to the left) it begins to climb significantly reveal something of how central that representation is to the instruction of a particular country on this topic.

Figure 5 reveals for France that four representations constituted something of a core of highly used representations. These included the symbolic representations of quotients, decimals, parts of a region and points on a number line. In terms of what is emphasized, the gray curves show that quotients and parts of a region were the most commonly emphasized representations in the core.

This core was supplemented by a "shell" of other interpretations, including all except fractions as ratios, which was virtually never emphasized and used basically only by those that reported making use of nine or ten representations. Of the others, fractions as comparisons, as measurements and as repeated unit fractions showed somewhat greater emphasis than did fractions as parts of a collection or as operators.

The core representations for New Zealand was similar to that for France, including the same four as before but in addition including relatively high use of fractions as repeated unit fractions and as parts of a collection. The level of emphasis for these later two representations suggested, however, that the core for New Zealand is not unlike that of France. The shell of supplementary representations was also very similar to that of France, except that slightly more use was made of fractions as ratios and less of fractions as operators.

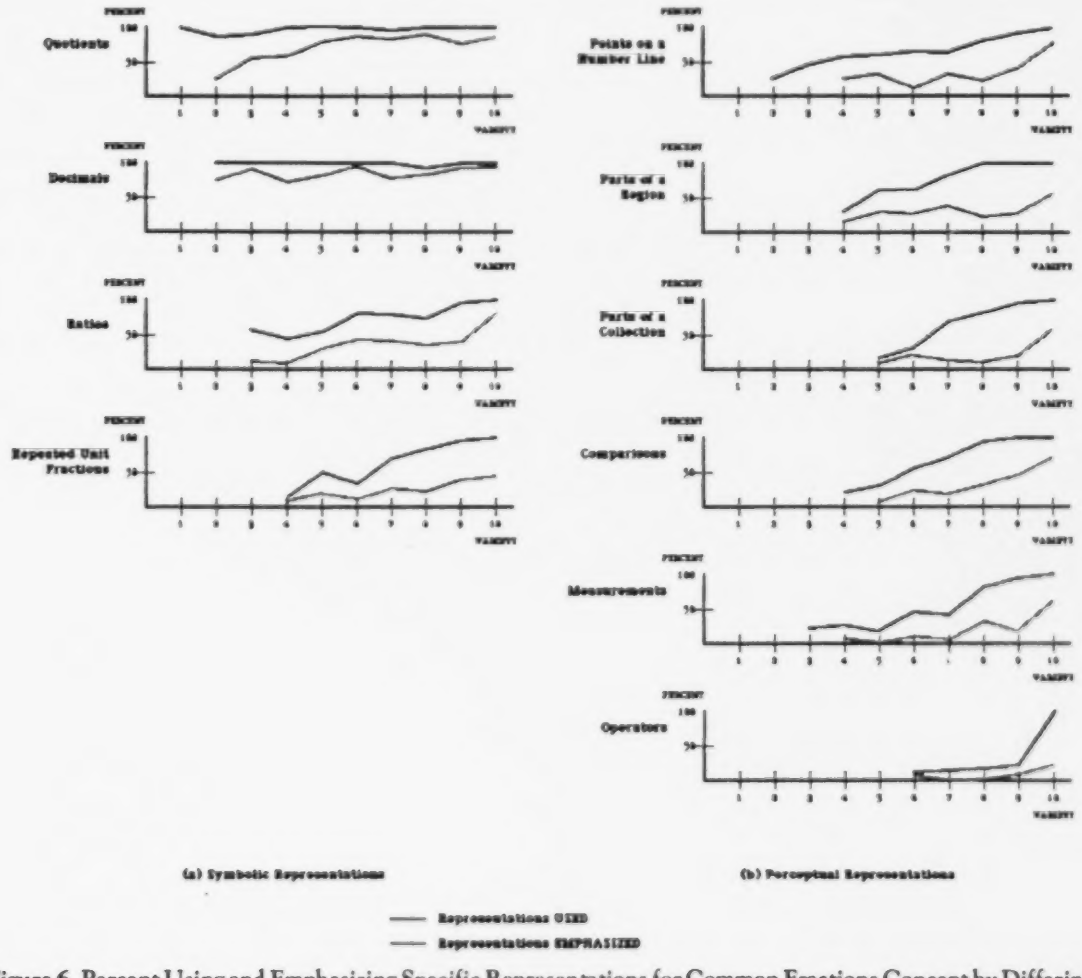
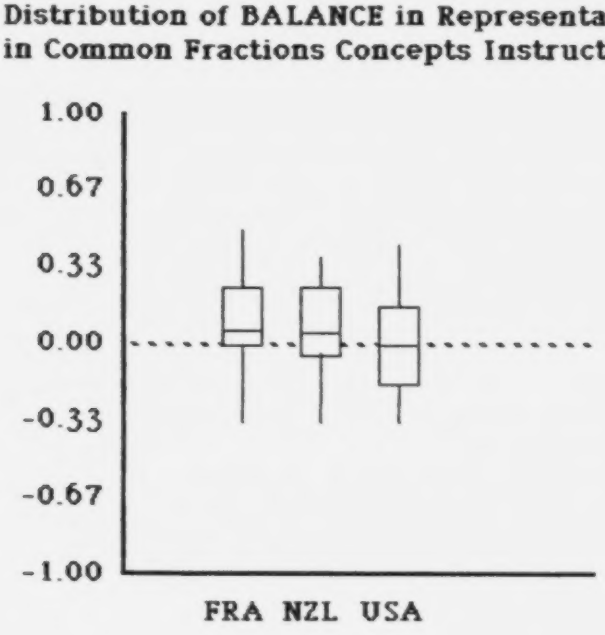


Figure 6. Percent Using and Emphasizing Specific Representations for Common Fractions Concept by Differing Variety for the United States.

Figure 6 shows that the core representations for the United States differed slightly from those of the other two. Core representations here included fractions as quotients , as decimals and as coordinates of points on a number line but a much less extensive use of fractions as parts of a region. Fractions as ratios received sufficient use and emphasis that it might well also be considered a core interpretation, in marked contrast to France and somewhat to New Zealand. Only fractions as operators appear not to be significant part of the shell of supplementary representations.

BALANCE, the other variable abstracted from the specific representations, offers some hope for being even more revealing and for having even more explanatory power than does the variable VARIETY.



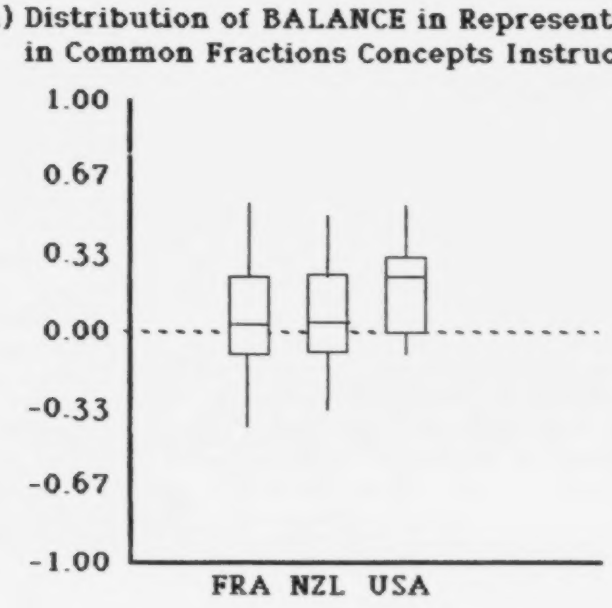


Figure 7. Distributions of Balance for Representations Emphasized or Used in Common Fractions Concept Instruction.

Figure 7 presents box-and-whisker plots of the distributions of BALANCE both for representations used (or emphasized) and for those emphasized only for instruction on the common fraction concept.

BALANCE scores greater than zero indicate relatively greater emphasis on the symbolic while scores less than zero indicate relatively greater emphasis on the perceptual. In Figure 7a the United States shows a distribution that is centered on zero and indicates a relative balance of emphasis in the representations used. By contrast, the other two countries show much more of an emphasis on the symbolic. Further investigation would be needed to determine just which representations provide that symbolic emphasis, but the distributions of the BALANCE variable are enough to reveal some clear national differences.

Figure 7b contains the distributions of BALANCE for just those representations emphasized and presents an interesting contrast to Figure 7a. The United States is seen to put relatively more emphasis on the symbolic than do the other two countries. Clearly there are differences between what is emphasized and what is merely used. This suggest that what is emphasized may have greater explanatory potential than consideration of what is used.

A similar picture emerged from examining BALANCE in fraction addition instruction. The United States again showed a relatively balanced use by a more symbolic emphasis. The pattern for France did not differ significantly from that in Figure 7. New Zealand both emphasized and used the perceptual more than the other two countries (or perhaps used the symbolic relatively less.

A Look at Effectiveness

This survey of the descriptive and explanatory potential of the SIMS data would not be complete without a look at the student achievement data and its links to the content representation data already discussed. Out of the pool of about 180 achievement test items at the Population A level of SIMS, 12 in particular dealt with common fractions concepts, computations and applications. Data for each of these twelve were examined and the patterns were similar regardless of whether the specific item dealt with concepts, computations or applications. A few basic points will be made here, but restriction to a single case study has limited explanatory findings to being suggestive at best.

The most obvious predictor of end of year performance on any item for any class is beginning of year performance on the same item. With this in mind, simple linear regressions were run for each item with classes from all three countries pooled.

The (Studentized) residuals for each class were plotted as an indication of whether that class, at end of year, performed above or below what might be expected based on its pretest performance. By looking at the set of residuals separately for each country, some indication of "overachieving" and "underachieving" countries can be seen.

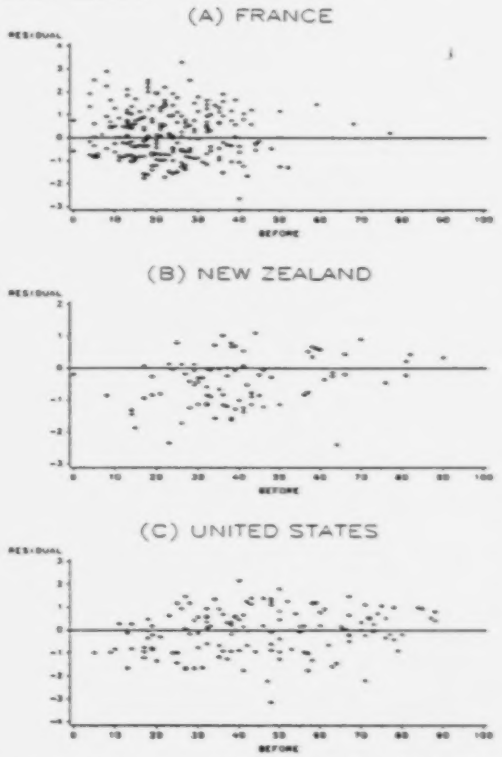


Figure 8. Pretest vs. Studentized Residual Class Achievement Scores for a Fraction as Part of a Regional Concept Item for Three Countries.

Figure 8 presents these plots for the common fractions concept item. Data for the other items were similar. It can be seen clearly that France had more classrooms with residual (gain) scores above zero than below; New Zealand had far more below than above; and the United States class residuals were scattered fairly evenly and randomly above and below zero. This indicates that, in comparison to the other two countries, France performed better than expected, New Zealand less well than might be expected and the United States somewhere in the middle. Results for the other achievement test items were similar. It should also be noted that the horizontal spread was different for France. Since common fractions were essentially new content for the target year in France, there were very few high pretest scores and thus much room for growth. This was not the case for the other countries.

The possible explanations for these outcomes are many. The explanation may be as simple as a

recency effect, since this material was new content in France and largely review content in the other two countries. More complicated explanations may be tied to the specifics of characteristic patterns of instruction or of teacher belief. Some of the more obvious analyses suggest that it may be hard to link achievement effectiveness to patterns of instructional strategy. For example, plots similar to those of Figure 8 but in this case the point for each class was marked to indicate whether the teacher emphasized, used, or did not use the representation of fractions as parts of a region (the most directly relevant representation) showed that in none of the three countries were high residual gains consistently associated with emphasizing that particular interpretation. The findings were similar for other items that could be linked to specific representations. Thus, emphasis of a particular representation could not be directly linked to high gains, even on achievement test items for which that representation was particularly salient.

Conclusions

An underlying theme of the work presented here has been to investigate the importance of specificity in description and explanation as opposed to the use of parsimonious and simplifying abstractions of the data (various variables and indices) which offer the potential for powerful explanation but at the cost of sacrificing concreteness and detail and running the risk of missed connections. As in almost everything else about the SIMS classroom process data, the results are mixed.

Certainly the specifics of description are rich and are worthy of further study for identifying important national characteristics. In contrast, in some cases the abstractions revealed patterns that were hard to see among the "trees" of the "forest". For instance, the VARIETY variable showed considerable difference between use of representation in conceptual and procedural aspects of common fraction instruction and the BALANCE variable showed some important characteristic patterns and some important differences based on what was emphasized versus what was merely used.

Neither strategy is adequate by itself for the search for effective description and explanation. The investigation of a large array of subtopics by more abstract indices has its place in exploring the presence or absence of characteristic patterns and general principles. However, such strategies must be supplemented by other investigations that focus in more detail on the specifics of a smaller number of cases to bring out characteristics and connections that might be missed otherwise.

TEACHING PRACTICES EMPLOYED IN THE TEACHING OF ALGEBRA AND GEOMETRY

David F. Robitaille

As part of the longitudinal component of the Second International mathematics Study, questionnaires were administered to participating teachers at the Population A (13-year old) level to obtain highly specific information about the teaching practices they employed in their classrooms. The five questionnaires, which were specially developed for use in the international study, dealt with the topics of algebra (integers, formulas, and equations); geometry; fractions; ratio, proportion, and percent; and measurement.

The importance of each of these topics in the Population A curriculum varies considerably from one jurisdiction to another, although algebra and geometry appear to be constant. That is to say, these two topics figure rather largely in the curriculum of each participating jurisdiction, although not equally so. By way of illustration, Table 1 presents a summary of the percent of class time in the Population A year devoted to the teaching of the five topics.

The results displayed in Table 1 for the teaching of algebra may be somewhat conservative estimates of the actual situation since, on that questionnaire, teachers were asked to report how much time they devoted to the teaching of integers, formulas, and equations only, and not to other algebraic topics which might form part of their curriculum. This means that in Belgium (Flemish) the teachers reported that they spent approximately 48 percent of the total time devoted to the teaching of mathematics in the Population A year to the teaching of integers, formulas, and equations. It may well be that additional time was spent dealing with other algebraic topics, but they were not asked about those on the questionnaire.

The caution expressed in the preceding paragraph applies to a certain degree to each questionnaire and to each jurisdiction in which the questionnaires were used. Although a questionnaire bears a certain content label, the precise connotation of that label is somewhat unclear. The applicability of the Algebra questionnaire to the French situation is illustrative.

In the curriculum analysis phase of the Second International Mathematics Study, France was categorized as being one of the jurisdictions which placed a heavy emphasis on the teaching of algebra at the Population A level; however, Table 1 indicates that French teachers stated that only 11 percent of class time was devoted to the study of topics covered in the Algebra questionnaire. This is undoubtedly a matter of the definition of the term "algebra"; i.e. what constitutes algebra in the French curriculum is probably different in many important respects from what constitutes algebra in the questionnaire developed for use in this study. That questionnaire dealt with the teaching of integers, formulas, and equations. Much of this material is treated in earlier grades in France and little or no time is devoted to it during the Population A year. We know from the questionnaire that French teachers spend about 11 percent of class time on the topics covered in the Algebra questionnaire. We do not know anything about how much time is spent on other algebraic topics.

Table 1
Time Spent on Questionnaire Topics
(Percent)

Topic	BFL	CBC	CON	FRA	JPN	NZE	THA	USA
Algebra	48	23	17	11	35	12	16	16
Geometry	27	17	13	37	17	15	12	12
Fractions	-*	16	14	20	-	12	14	17
Ratio, Prop., Percent	-	11	12	6	-	5	8	11
Measurement	-	12	14	3	-	8	9	8
TOTAL	75	79	70	77	52	52	59	64

*Questionnaire not used.
BFL = Belgium (Flemish), CBC = Canada (British Columbia), CON = Canada (Ontario), FRA = France, JPN = Japan, NZE = New Zealand, THA = Thailand, USA = United States of America.

Bearing in mind the limited scope of the curricular content covered by the questionnaires and the inherent limitations of self-report data, it is important to recognize the uniqueness and importance of their contribution to our knowledge about what transpires in mathematics all around the world. The questionnaires were designed especially for use in this study, and were extensively pilot-tested in several of the participating jurisdictions to enhance the validity of the results obtained. Little is known about what actually transpires in classrooms, and these questionnaires provided a way of obtaining comparative data from a variety of jurisdictions on the teaching practices employed in the teaching of mathematics.

Of the five questionnaires, two were used in all eight participating jurisdictions. In some places only two were used because the topics treated in the other questionnaires were not as important in the mathematics curriculum at that level; in others, it was decided that asking teachers to complete five extensive questionnaires was not a good idea. In this paper, results from the two questionnaires used in all eight systems are considered. An analysis of the data from all five questionnaires will form part of the international report of the longitudinal component of the study. that report is expected to appear in the near future.

The Teaching of Algebra

The box-and-whisker (Tukey, 1977) plots in Figure 1 summarize the distributions of amounts of time spent on the topics covered in the Algebra questionnaire. The median number of hours internationally was 23. Belgium not only reported the highest median number of hours spent on algebra, 67 out of a total of 140, but it also had the widest spread, indicating a considerable degree of variation within the country. All of the other countries have fairly narrow spreads.

The graphs for all but two of the systems include several outliers, especially those for France and the United States. For the United States, the outliers represent Population A classes taking a full year of algebra, while most American students would not take such a course until the year following the Population A year.

Topics Taught

Eleven topics under the sub-headings Integers, Formulae, and Equations, were treated in the Algebra questionnaire. Taken together these eleven topics constitute the definition of algebra at the Population A level for the classroom-process phase of the study. A list of the topics and the percent of teachers who either taught or reviewed them is shown in Table 2.

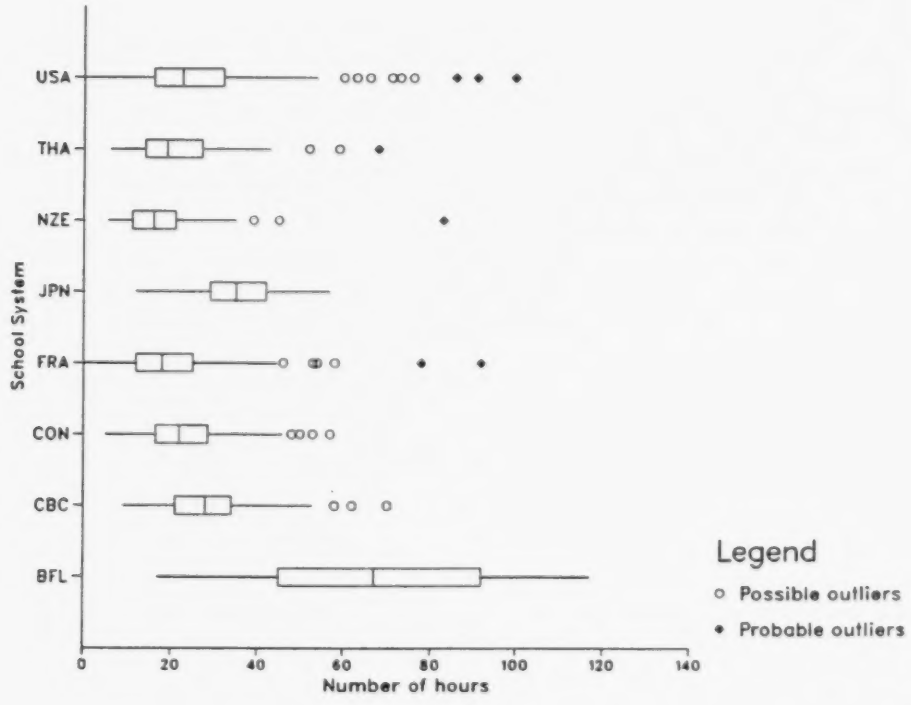


Figure 1. Time spent on algebra

Table 2
Algebra Topics Taught or Reviewed
(Percent)

Topic	BFL	CBC	CON	FRA	JPN	NZE	THA	USA
Integers								
-concept of	76	100	95	73	99	98	100	99
-addition	87	100	95	87	99	98	100	99
-subtraction	89	100	95	88	99	98	100	98
-multiplication	85	100	92	89	99	95	98	98
-division	86	100	91	91	99	93	99	97
-properties	89	85	65	90	98	75	99	82
-order relations	81	92	82	93	99	96	92	93
evaluate formulae	75	95	95	82	99	90	96	94
derive formulae	42	64	62	60	99	31	82	61
solve literal eqns.[1]	40	30	42	50	68	19	79	40
solve linear eqns.[2]	95	96	92	100	99	87	97	92

[1] linear equations of the first degree, in one unknown, with literal coefficients
[2] linear equations of the first degree, in one unknown, with numerical coefficients.

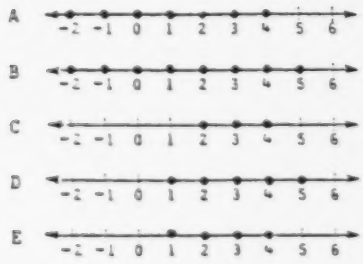
Of the 11 topics, nine were either taught or reviewed by virtually all teachers in every country. The exceptions were deriving formulae or equations and solving literal equations. These were taught by significantly fewer teachers than the other topics.

The major differences among systems regarding coverage of the 11 topics was whether the material was considered to be new for students at this level or was customarily taught earlier. In France and Belgium, almost all of the material dealing with integers is apparently taught before the Population A year and reviewed or extended during this year. In Japan, on the other hand, almost all teachers reported that the material dealing with integers was being presented to students as new material. There was considerably less unanimity within most countries regarding the teaching of topics dealing with formulae and equations. Only the Japanese had high proportions of teachers indicating that these topics were taught to students for the first time in the Population A year.

Integer Concepts

The choice of which pedagogical approaches to use in the teaching of algebra seems to depend on the subject matter being presented and whether or not the topic is being introduced for the first time. For example, in introducing the concept of an integer, over 70 percent of Population A teachers in countries other than Belgium and France emphasize the use of a number line where the integers are seen as an extension of the natural numbers, or as coordinates of points on the number line. The number line may also be used to illustrate operations with integers, particularly addi-

The set of integers less than 5 is represented on one of the number lines shown below. Which one?



Item 082

	Pretest (%)			Posttest (%)			Change
	Correct	Incorrect	Omit	Correct	Incorrect	Omit	
BFL	48	47	5	54	43	3	6
CBC	-*	-	-	66	28	6	n/a
CON	37	55	8	56	42	2	20
FRA	46	28	26	58	34	8	12
JPN	40	57	2	55	43	2	15
NZE	41	54	4	57	42	1	16
THA	26	69	5	39	60	0	13
USA	39	56	4	51	49	1	11

*Item not included in the pretest.

tion, subtraction, and multiplication.

Item 082 was the only test item that dealt directly with the use of the number line as a means of representing integers. Growth scores exceeded 10 percentage points in six of the seven systems that made use of this item in the pretest, but pretest scores were quite low. Posttest scores were rather disappointing, with the highest being 66 percent correct in Canada (B.C.). In every other place, the posttest score was less than 60 percent.

It is difficult to explain why students seemed to find this item so difficult. There was universal agreement that the item was appropriate for this level, and that the material had been taught. Incorrect answer choices were divided more or less equally among the four distractors, and rates of omission were not at all high.

Item 014 also involved use of the number line, but in that instance students were asked to order three numbers including a negative rational: viz. – 1/2. Overall, performances was no better on this item than on Item 082.

Over 70 percent of teachers in every country except Belgium and France reported that they emphasized the use of the number line in teaching integer concepts. In France and Belgium, where this material is usually taught initially before the Population-A year, the approach taken is much more abstract and related to axiomatic structures. Thus, Belgian teachers were much more likely to refer to integers as vectors or directed segments than teachers elsewhere, while more teachers in France than anywhere else emphasized a definition of an integer as an equivalence class of ordered pairs of whole numbers:

$$-2 = \{(0, 2), (1, 3), (2, 4), \ldots\}$$

or

$$-2 = \{(a, b) \in W \times W \mid b = a + 2\}$$

Another approach which is commonly used in the teaching of integers, everywhere except in Thailand, is the employment of examples of physical situations involving integers. Students discuss situations such as heights above or below sea level, temperatures above and below zero, and profit and loss in which integers are used as vector quantities to convey a sense of both quantity and direction. Such examples were reported as being emphasized particularly by teachers in Canada (Ontario) and Japan where integer concepts are introduced for the first time at this level and which had the youngest students participating.

Achievement results on items related to real-world applications of integers were rather disappointing. For example, on Item 013 students were asked to tell how much warmer a temperature of 31 degrees was than one of –7 degrees. The highest posttest score on this item was 70 percent in Belgium (Flemish). Next was Japan at 63 percent and all the rest were less than 60 percent. The most popular distractor by far was 24 degrees, the algebraic sum of 31 and (– 7). Over 20 percent of students in each of the eight systems chose this response. Given such relatively poor posttest results, it is not at all surprising to find that growth scores were very low: the highest was 10 percentage points in each of the two Canadian provinces. Teachers everywhere considered the item to be an appropriate one, and indicated that students had been taught the concepts and techniques involved. In spite of this, posttest results were quite poor.

Operations with Integers

Whether or not operations with integers such as addition, subtraction, and multiplication are being taught for the first time, teachers in most countries say that they emphasize rules for performing those operations rather than other approaches which attempt to build meaningful rationales for the algorithms employed. Exceptions to this trend were reported primarily in Canada and New Zealand.

Teachers in all countries are strongly of the opinion that students require a great deal of practice in order to become proficient in performing operations with integers. They also believe that students are not very interested in knowing why rules for performing operations with integers work, and this opinion undoubtedly contributes to their emphasis on such rules.

Performance on the three test items dealing with operations with integers (Items 012, 049 and 113) resulted in much greater growth scores overall, and higher posttest scores than was the case for items dealing with integer concepts. For example, on Item 012 which required students to find the product of (–2) and (–3), the performance of students in Thailand increased by 53 percentage points, and in Ontario by 47 points between pre- and posttest. In Japan, on Item 113, which required students to find the difference

$$(-6) - (8)$$

performance grew by 53 points, to 72 percent correct. However, only one national posttest score on any of these computational items exceeded 80 percent, and there is some reason to doubt that students had achieved mastery of these algorithms in spite of the opinions expressed by their teachers about the importance of practice.

Item 012

(–2) x (–3) is equal to:

A. –6 B. –5 C. –1 D. 5 E. 6

	Pretest (%)			Posttest (%)			Change
	Correct	Incorrect	Omit	Correct	Incorrect	Omit	
BFL	66	33	1	78	21	2	12
CBC	36	59	5	72	27	2	36
CON	14	83	3	60	39	1	47
FRA	72	27	2	79	20	1	7
JPN	-	-	-	85	15	0	n/a
NZE	13	86	1	47	52	1	34
THA	9	91	1	62	38	0	53
USA	24	74	2	56	44	0	32

Item 049

–5 (6–4) is equal to:

A. 50 B. 26 C.10 D. –10 E. –26

	Pretest (%)			Posttest (%)			Change
	Correct	Incorrect	Omit	Correct	Incorrect	Omit	
BFL	68	24	8	75	22	2	7
CBC	–	–	–	75	18	7	n/a
CON	58	31	11	65	32	3	7
FRA	66	24	10	75	19	5	10
JPN	–	–	–	78	21	1	n/a
NZE	53	38	9	61	36	2	8
THA	57	39	4	59	40	1	1
USA	59	35	6	65	34	1	5

Item 113

(–6) – (–8) is equal to:

A. 14 B. 2 C. –2 D. –10 E. –14

	Pretest (%)			Posttest (%)			Change
	Correct	Incorrect	Omit	Correct	Incorrect	Omit	
BFL	46	52	2	57	42	1	11
CBC	-	-	-	49	49	2	n/a
CON	16	81	3	43	55	1	27
FRA	70	28	1	70	29	1	0
JPN	19	74	7	72	27	1	53
NZE	19	79	2	30	69	1	11
THA	17	82	1	32	68	0	15
USA	24	73	3	41	58	1	17

Solving Equations

In teaching students how to solve equations of the first degree in one variable, e.g.

$$7x + 5 = 40$$

teachers in all countries reported emphasizing an algebraic approach based either on properties of equality or on the properties of additive and multiplicative inverses. Few emphasized other possible techniques such as trial and error. While this is perhaps not a surprising finding, it underscores an apparent tendency among teachers at this level to stress formal mathematical approaches to topics rather than more intuitive ones. It is particularly interesting to note that this is a widespread, if not a universal, tendency.

Two test items dealt explicitly with the solution of equations. On Item 086, students were required to solve the equation $\frac{4x}{12} = 0$; on Item 151, $5x + 4 = 4x - 31$. On neither item were there any posttest scores greater than 60 percent, and even in cases where growth was substantial the overall results were disappointing. For example, scores on Item 151 increased by 21 and 24 percentage points in Belgium and France, respectively. However, their posttest scores were only 53 and 42 percent correct. This can hardly be interpreted as a positive result.

Summary

The general impression that one obtains from studying performance on the algebra test items is that students found them difficult. Posttest scores were generally low, often surprisingly so. Teachers report having taught this material and they appear to emphasize rules and abstract justifications in their teaching. These results point out a need for teachers, researchers, and curriculum developers to re-examine the teaching of introductory algebraic concepts and techniques to see whether this material can be taught more successfully at this level, or perhaps to recommend that these topics be delayed until students are better prepared to assimilate them.

The Teaching of Geometry

The box-and-whisker plots shown in Figure 2 summarize the number of hours devoted to the study of geometry at the Population A level. Students in France spend twice, and in some case three times, as much time on geometry as students in most other countries. In Belgium the median number of class hours per year for geometry was slightly lower than in France: 37 out of a total of 140 hours of mathematics for the year.

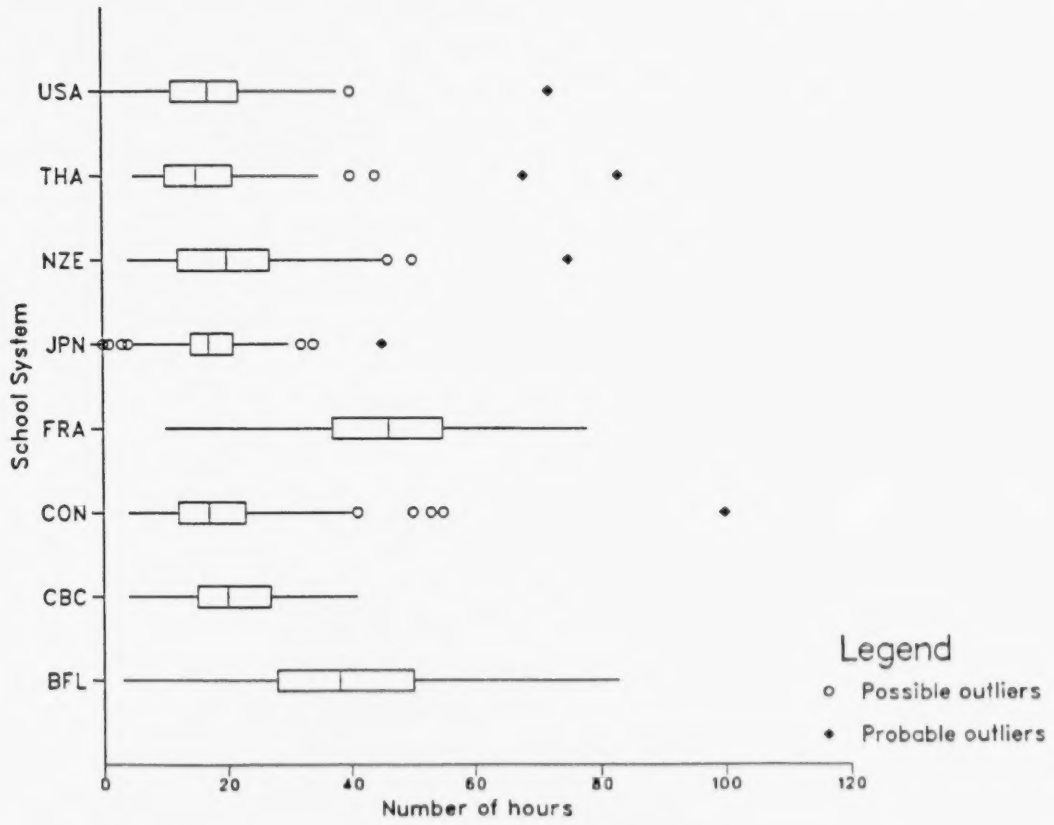


Figure 2. Time spent on geometry

In the other participating systems, less than 20 percent of class time during the year appears to be devoted to geometry; in several cases, considerably less. The specific results are as follows:

Canada (B.C.)	15%
Canada (Ontario)	10%
Japan	17%
New Zealand	12%
Thailand	12%
United States	8%

In most cases, the amount of time devoted to geometry was significantly less than that devoted to algebra. In the United States and Canada (Ontario), teachers reported spending more time on fractions than on geometry, and such results should be a cause for concern to mathematics educators, curriculum developers, and classroom teachers.

Problems with the status of geometry in the mathematics curriculum have been apparent for years. Since at least the time of the Royaumont Conference in 1959 (OEEC, 1961) and Dieudonné's ultimatum to the effect that "Euclid must go!", the geometry curriculum has been in disarray. What these data may indicate is that teachers, faced with such disarray, have decided that their valuable and limited class time would be better spent working on areas of the mathematics curriculum other than geometry. Geometry, according to these data, may well be on the endangered species list in mathematics education.

Geometric Content in the Curriculum

Sixteen topics, ranging from highly specific ones such as the Pythagorean Theorem to fairly broad themes such as transformations, were listed on the geometry questionnaire. Teachers were asked to indicate whether or not each of these topics formed part of their geometry curriculum, and whether the topics they did teach were taught as new or as review material. The sixteen topics were:

angles (right, acute, supplementary, etc.)
transformations (translations, reflections, rotations)
vectors
the Pythagorean Theorem
triangles and their properties (excluding congruence)
polygons and their properties (excluding properties related to congruent or similar polygons)
circles and their properties
congruence of geometric figures (including triangles)
similarity of geometric figures (including triangles)
parallel lines
spatial relations
geometric solids and their properties
geometric constructions with ruler and compass
proof (formal deductive demonstrations)
tessellations
coordinate geometry

Treatment of these topics varied considerably in different systems, as is shown by the median polish results in Table 3. Positive entries in the table indicate that a particular topic is given comparatively more importance in a given school system; negative entries indicate the opposite. Results greater than 15 in absolute value were considered significant for this analysis, and they are printed in bold type in the table.

The large positive values in the rightmost column correspond to topics that are most likely to have been taught in these systems. The six topics so indicated are typical selections from plane Euclidean geometry: angles, triangles, polygons, circles, parallel lines, and ruler-and-compass constructions. The three topics with the most negative weightings — vectors, spatial relations, and proof— are the least likely to be taught among the sixteen listed.

This set of topics did not fit the curriculum particularly well in Belgium or France. The individual cell residuals for those countries show that they place much greater emphasis on transformations, vectors, and formal proof than do teachers elsewhere. These two countries also show significant negative residuals for many of the Euclidean topics, indicating that these topics are not given much importance at the Population A level. In fact, except for the topic "angles", Belgian and French teachers reported that many of those topics did not form part of their geometry curriculum prior to the Population A level either.

Teaching Practices Employed

These curricular disparities are reflected in the achievement results. Consider, for example, Item 122 shown below. The item deals with the sum of the angles in a triangle, and is a typical item of the kind included in an introductory treatment of Euclidean geometry at this level.

Posttest performance on this item was very high in Japan at 89 percent, and fairly good in Canada (B.C. and Ontario), New Zealand, and Thailand, where almost all teachers reported teaching this topic. Substantial growth was also reported in Canada and New Zealand. Students in Belgium (Flemish) and France did less well: 63 percent and 55 percent, respectively. In these two places, almost half the teachers indicated that this material had not been taught.

In the United States, where fairly strict streaming

of students into different mathematics courses is widely practised at this level, posttest achievement on the item was low and only half the teachers reported teaching the material. In other words, although the United States results were very similar to those from Belgium and France on this item, the factors underlying those performance levels were very different.

Achievement levels on the four items dealing with transformational geometry in a fairly formal way were quite poor, even in Belgium and France, where a transformational approach is emphasized. For example on Item 173, shown below, the highest posttest score was only 20 percent correct. Students in France and Belgium seemed to find these items as difficult as students elsewhere did, in spite of their reported emphasis in the curricula of those countries.

When these data are combined with a description of the basic instructional approach to geometry taken by teachers, yet another indication of the disparity that exists among countries with respect to the geometry curriculum becomes clear. Teachers in Belgium, France, New Zealand and Thailand favor a transformational approach. In New Zealand, the approach is characterized as an informal one, whereas it is much more formal in the other three. North American teachers are much more likely to use an informal Euclidean or coordinate approach to geometry, and not to stress formal proof at all. In Japan, the approach is Euclidean, but there is some ambivalence about the degree of rigor used.

There is also some apparent ambivalence in the opinions expressed by teachers in certain countries regarding the best way to teach geometry at this level. For example, in spite of the relatively formal nature of their instructional approach and curriculum, about 60 percent of Belgian, French and Thai teachers agreed that, "An intuitive approach to geometry is more meaningful to students at this grade level than a formal approach." Moreover, although a majority of teachers in these three countries agreed that it was desirable to follow an axiomatic approach, there was not a strong consensus of opinion to that effect.

The Role of Proof in the Geometry Curriculum

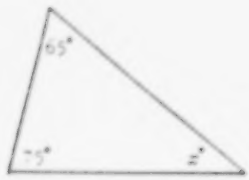
A clear difference of opinion exists on the appropriateness of proving theorems for students of this age. Teachers from Canada, New Zealand, and the United States are much more likely to agree that such activity should be postponed to a later grade when students are older, and presumably, more mature. Teachers from the other countries, and particularly those from France, are much less likely to agree with that opinion. Students' achievement levels on items involving proof in geometry were quite low in all countries, and it seems evident that students at this age level find such reasoning difficult.

Table 3
Geometry Topics Taught or Reviewed
(Median Polish)

	BFL	CBC	CON	FRA	JPN	NZE	THA	USA	ROW EFFECTS
Angles	-16	0	0	-60	24	13	-12	0	31
Transforms.	61	-41	-4	39	10	41	-21	-36	-1
Vectors	112	-30	-8	66	-2	11	4	-21	-29
Pyth. Thm.	0	33	5	-51	-21	-12	33	17	-7
Triangles	-30	3	-2	0	-17	10	0	-4	30
Polygons	-15	1	-1	8	-10	9	-20	0	20
Circles	-28	-4	4	11	7	-44	-26	2	21
Congruence	-14	12	13	-20	-20	-11	28	12	7
Similarity	-4	5	13	-51	-26	-10	34	13	0
Parallel Lines	40	-7	-4	7	-6	0	-1	-4	33
Spatial Rel'n.	14	-6	2	-12	72	0	8	-7	-19
Solids	0	-4	4	-14	54	-13	24	8	-2
Const.	0	0	4	17	32	-19	-16	-20	23
Proof	102	-26	-5	64	11	-10	51	-21	-31
Coordinates	6	5	-6	15	0	26	-13	10	-14
Column Effects	-50	8	6	0	-25	-4	6	0	60

Item 122

	Pretest (%)			Posttest (%)			Change
	Correct	Incorrect	Omit	Correct	Incorrect	Omit	
BFL	61	31	7	63	33	5	1
CBC	47	38	15	73	22	6	26
CON	53	41	5	72	26	2	19
FRA	49	31	20	55	30	15	6
JPN	-	-	-	89	10	1	n/a
NZE	58	41	1	75	25	0	17
THA	65	34	1	72	28	0	7
USA	37	58	5	56	42	1	19



x is equal to

A 75

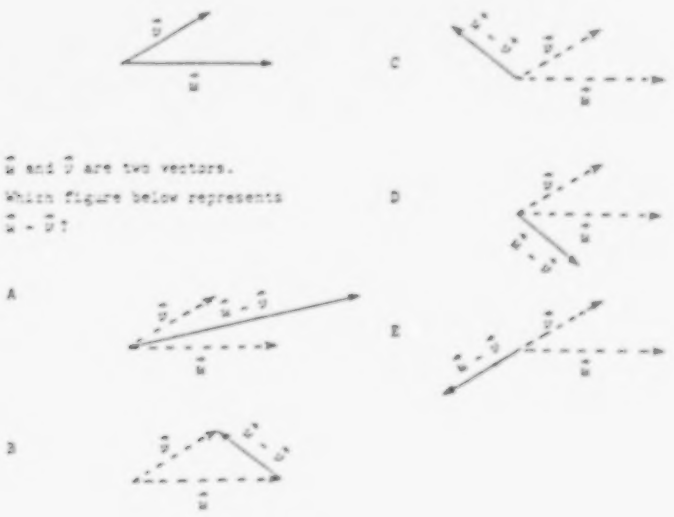
B 70

C 65

D 60

E 40

Item 122



Item 173

Summary

There is a remarkable degree of consistency among the teachers who participated in this study regarding the methods and materials to be used in the teaching of mathematics and in their opinions about issues in mathematics education. As an example of the latter, teachers repeatedly and universally disagreed with all of the statements on the questionnaires which suggested that calculators should be used extensively in mathematics classes at this level.

There is also an apparent consensus among teachers in all of these countries, except France, that students need to be taught the same material over and over again. No matter what the topic, very few teachers said that they took it for granted that students had encountered and mastered this material in an earlier grade or grades. Only teachers in France reported doing so with any degree of frequency.

The implications of such a practice for the teaching of mathematics are enormous. If teachers believe that they cannot assume that students have mastered and retained material which they have seen in previous grades, then a tremendous amount of reviewing must be done. Such a practice would seem to be wasteful in terms of the amount of time consumed, and stultifying for students who have to work through the same material over and over each year.

Previous studies of teaching practices conducted in North America have concluded that the teaching of mathematics is largely a teacher-directed, "chalk-and-talk" affair (Romberg and Carpenter, 1986). The results of this study add further confirmation to this conclusion. There are many instances in the data where teachers indicated agreement with a statement, but reported doing exactly the opposite in practice. For example, they agree that having students measure and explore all important activities in teaching of measure-

ment, but they also say that they do not put this opinion into practice.

The reasons for this lack of congruence between opinion and practice are unclear. It may be that teachers find themselves so pressed for time to complete the prescribed curriculum that they cannot afford to devote any extra time to laboratory-like approaches. Or, it may be that they are unwilling to do so.

References

Organisation for European Economic Co-operation, (1961) *New Thinking in School Mathematics*. Paris: O.E.E.C. Publications.

Romberg, T.A. and Carpenter, T.P. (1986) Research on teaching and learning mathematics: two disciplines of scientific inquiry. In M.C. Wittrock (Ed.) *Handbook of Research on Teaching* (Third Edition). New York: Macmillan Publishing Company.

Tukey, J.W. (1977) *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley Publishing Company.

IDENTIFICATION AND DESCRIPTION OF OPPORTUNITY TO LEARN AND GROWTH IN ACHIEVEMENT

Richard G. Wolfe

Mathematics is a topic that is mostly learned in school, so the context for assessing mathematics achievement needs to be the teaching and learning environment of the mathematics classroom. The IEA Second International Mathematics Study (SIMS) looked at mathematics achievement and its environment from the perspectives of:

1. the intended curriculum, defined by national and local syllabuses, guidelines, and regulations, by the contents of textbooks, and by school structures including tracking and retention;

2. the implemented curriculum, defined by teachers' reports of their individual goals and attitudes, of their use of instructional resources, of their teaching methods, and, especially, of the actual time spent and specific mathematical material covered; and

3. the attained curriculum, defined in terms of what mathematics knowledge students acquire and also their attitudes toward mathematics and mathematical study.

The SIMS survey was carried out in the early 1980's in some twenty countries. Two levels of school mathematics were studied: Population A, corresponding to the grade in which the modal student age was 13 years, and Population B, corresponding to students specializing in mathematics in their final year of secondary school. The surveys included extensive background, attitude, and pedagogical questionnaires for teachers, school principals, and students in addition to student achievement testing. The SIMS is partly a replication of an earlier international study, described by Husén (1967), that was carried out in the early 1960's in twelve countries.

In most developed countries, Population A is the last level of schooling where education, and particularly mathematics education, is essentially universal: most 13-year-old children are still in school and still taking mathematics. There are, however, important differences within and between countries in what mathematics is taught and how it is taught. In some contexts there is repetition of earlier instruction in arithmetic. In other contexts, there is introduction of new topics, especially algebra and geometry. There is variation in the extent of abstraction and symbolism used in presenting mathematical ideas.

This paper focuses on Population A results obtained within SIMS for eight "countries" (Flemish Belgium, British Columbia, Ontario, France, Japan, New Zealand, Thailand, and the United States of America.) that used the full methodological design of the SIMS, including:

1. longitudinal achievement testing: the students were pretested at beginning of school year and posttested at the end of the school year, using a pool of 176 or 180 mathematics items (through a test form rotation scheme, not all students had to answer all items);

2. opportunity-to-learn measurement: the teachers of the sampled classrooms indicated for each test item in the pool whether their students had the opportunity to learn the mathematics necessary to give a correct answer; and

3. classroom process description: special questionnaires were filled out by the teachers during the school year to provide rich description of classroom processes, concerning both methods for teaching specific mathematics topics and general pedagogical styles.

These are important methodological innovations in large-scale, international educational surveys (or for that matter, for national or local studies) and allow detailed description of what is taught and learned in one year, disentangling that from cumulative knowledge gained over a student's school career. It is also possible to make correctly specified correlation of within-year learning with within-year classroom characteristics and processes.

In Japan, nearly all the teachers (93 percent) consider the item to be old content, and while the students perform rather well on the item, there is no growth over the school year: in the pretest 63 percent get the item

Issues concerning the research design

The array of data for the SIMS longitudinal, classroom process Population A study is shown in Figure 1.

In each country, a complex sample was drawn, starting with basic stratification of schools according to jurisdictional or geographical categories. The general pattern was then to sample schools within stratum with probabilities proportional to size or estimated size, to sample two classrooms at random from each sampled school, and then to regard as the final sampled units the teacher and all the students of the selected classrooms. The final sizes of the sample varied by country from 93 to 365 classrooms and 2567 to 8778 students. The basic survey statistics—viz., the percentage of correct item response—have standard errors of 1 or 2 percent, as estimated from the variability of classroom and school means.

The research design is discussed fully in Burstein (1988). For the purposes of this paper, we need to consider the critical issue of the definition of the cognitive achievement measures.

In an international educational survey, the achievement tests are inevitably compromises, because national curricula vary significantly in content and em-

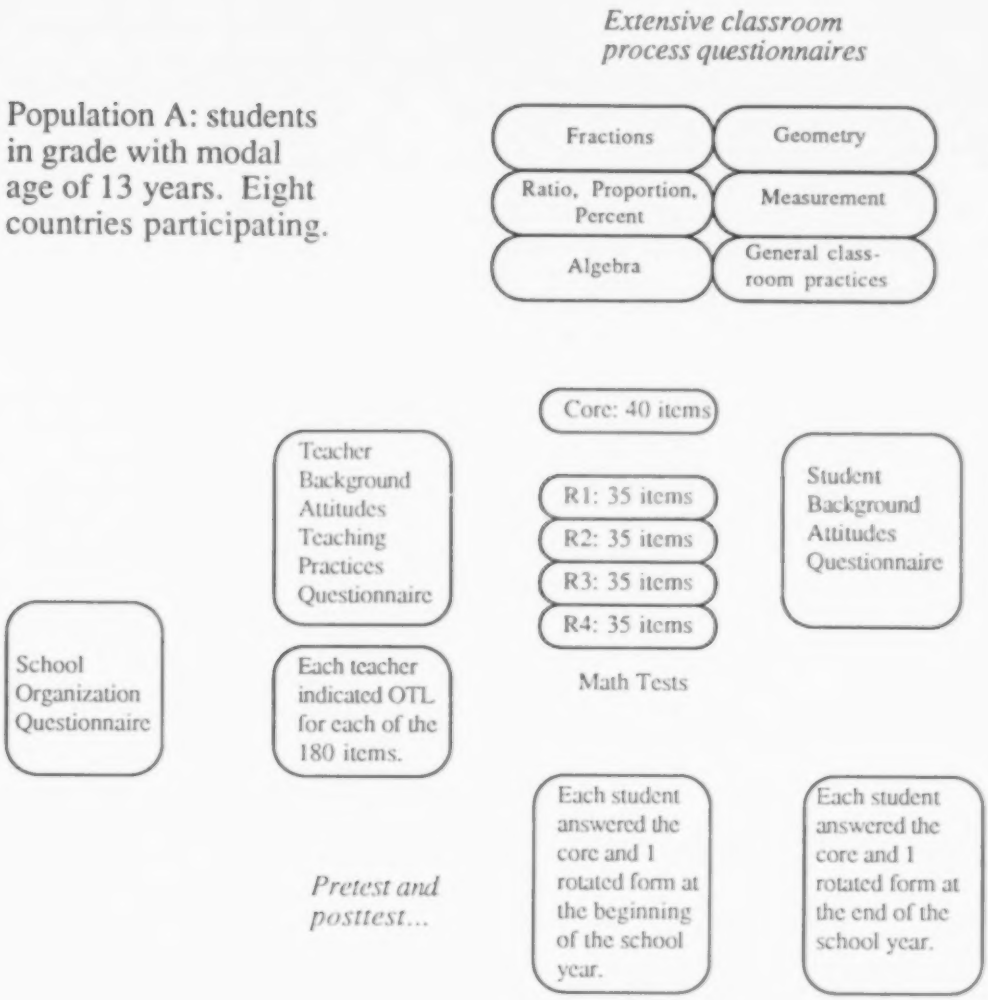

Extensive classroom process questionnaires
Population A: students in grade with modal age of 13 years. Eight countries participating.
Fractions
Geometry
Ratio, Proportion, Percent
Measurement
Algebra
General classroom practices
Core: 40 items
R1: 35 items
R2: 35 items
R3: 35 items
R4: 35 items
Math Tests
Teacher Background Attitudes Teaching Practices Questionnaire
Student Background Attitudes Questionnaire
School Organization Questionnaire
Each teacher indicated OTL for each of the 180 items.
Pretest and posttest...
Each student answered the core and 1 rotated form at the beginning of the school year.
Each student answered the core and 1 rotated form at the end of the school year.

Figure 1. Data array for the longitudinal, classroom process component of the IEA Second International Mathematics Study.

phasis. In some countries, notably the United States of America, there is substantial curriculum variation within country. An initial focus of the SIMS project was to describe and report the curriculum variation, and a book of analysis has been prepared (Travers & Westbury, 1989). The mathematics item pools and classifications were derived from the analysis, and the final selection of items for the international testing were determined by ensuring that for each country, its most important Population A contents were included, and that over all countries, more items were used for content areas that were important in a majority of countries. This works reasonably well in that there are some contents that are taught most everywhere: basic arithmetic including fractions, the concepts of integers, methods for handling ratio, proportion and percent, and some beginning algebra. On the other hand, there are topics that are not taught in some countries: for example, square root is not taught at this level in Japan. And some topics are taught with special content and emphasis: for example, in France and Belgium, geometry is taught from a formal, transformational perspective.

Such differences can make international comparisons in achievement quite misleading, unless the comparisons are made for specific content areas and are considered relative to degrees of national emphasis and opportunity to learn. And the determination of specific achievements means that the mathematics knowledge domain must be finely articulated and that there need to be many mathematics test items employed.

Cognitive Response and Opportunity to Learn

The basic findings of the longitudinal SIMS survey are to be found in the item-by-item tabulations of cognitive response and opportunity to learn. An example for an item in the Ratio-Proportion-Percent topic is given in Figure 2. For Japan, this item was part of the pretest given all students and was on a rotated form for the posttest, so it was given to 25 percent of the sample. In the other countries it was on the core test and so was taken by all students both at pretest and posttest.

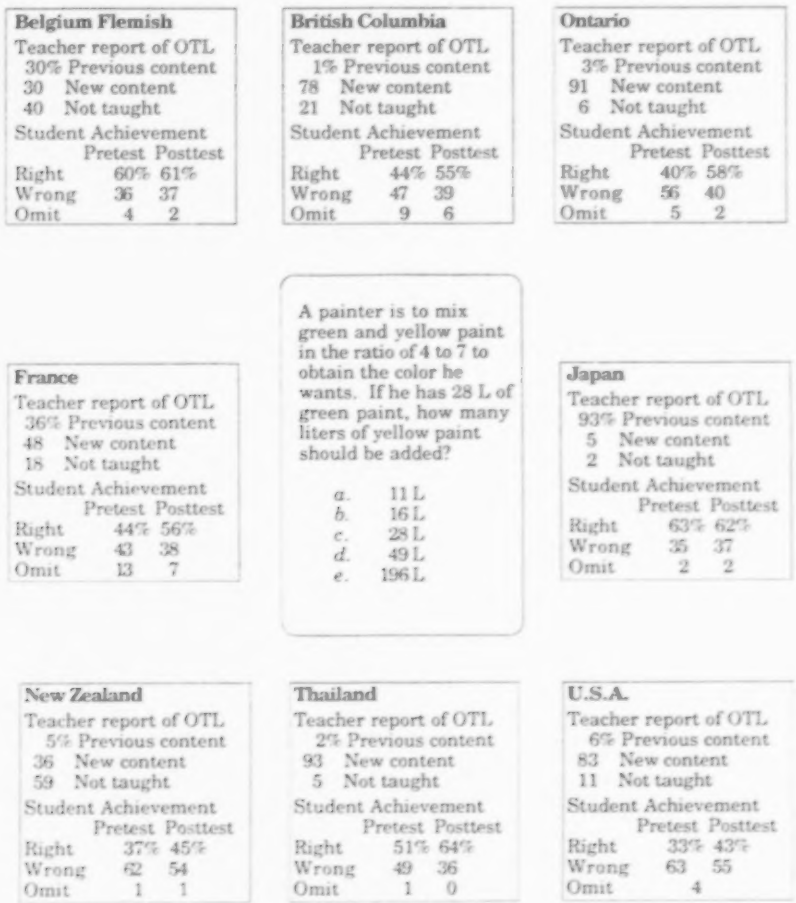


Figure 2. Opportunity to learn and pretest and posttest achievement across countries on one ratio question.

In Thailand and Ontario, we see the opposite circumstance. In Thailand, nearly all the teachers (93 percent) consider the item to represent new content that was taught during the year, and there is student cognitive growth from 51 percent correct in the pretest to 64 percent correct in the posttest. Similarly in Ontario, 91 percent of the teachers taught the mathematics for this item as new material, and the students showed growth from 40 percent correct in the pretest to 58 percent correct in the posttest.

The United States of America., British Columbia, and New Zealand show few teachers who regard this item to represent old content (6 percent, 1 percent, and 5 percent) and progressively decreasing percents of opportunity to learn this item as new material (83 percent 78 percent, and 36 percent). The student achievements and levels of cognitive growth are correspondingly low: 33 percent, 44 percent, and 37 percent at the pretest going to 43 percent, 55 percent, and 45 percent at the posttest.

The results are more confusing for Belgium Flemish and France, because some teachers regard the content to be old and others regard the content to be new. The students in Flemish Belgium perform well on the item but show no growth (60 percent correct on the pretest, 61 percent on the posttest); the teachers seem to be split evenly between regarding the item as old content, as new content taught, or as content not taught. In France, nearly half (48 percent) of the teachers report teaching the mathematics for the item, but another 36 percent regard the item's content as old, and the student show some growth, from 45 percent correct the pretest to 56 percent correct at the posttest.

Informal Transformations in Geometry

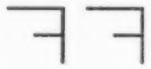
All of the mathematics testing in SIMS was done within a five-alternative, multiple-choice format. While the validity of the interpretation of the item response and its correlates depends primarily on the logical and empirical connections made between the mathematics test item and the mathematics curriculum—intended and implemented—the interpretation also hinges on an understanding of the students' response processes, which are as much psychological as mathematical. The parameters of the processes may be affected by and change during the year of instruction. The multiple-choice response mode imposes inherent limitations on how much one can tell about how a student responds.

In particular, in making international comparisons, one must consider how the item response patterns vary between countries. A major point of difference is the tendency for students in some countries to omit responding when they are evidently unsure of their knowledge contrasted to the behavior of students in other countries to try to answer each question—perhaps by "guessing". The international instructions did not advise students to guess nor threaten any "correction" in the scoring, but simply stated that these were international tests and that some items would be unfamiliar to them.

The students in France were inclined to omit responses, with the omission rate approaching 50 percent for some items. According to the French study director, students in France are expected to be able to defend their answers: guessing would not be considered appropriate. The omitting rate in Thailand is, on the other hand, less that 1 percent for most items. A more detailed analysis of the Thai data has shown little correlation between wrong responses at the beginning of the year and wrong responses at the end of the year: that is, students must feel obliged to answer each question and are guessing when they do not know the answer. For the countries with omission rates between these extremes, there is some evidence for systematic misinformation (viz., same wrong response at the beginning and end of the year) and some evidence for seemingly random responses. But there is no justification for a general "correction" for "guessing" adjustment to the response data.

One way to handle the omitting-guessing ambiguity is to preserve, throughout the interpretation, a three-way tabulation of item responses, considering rights, wrongs, and omits at pretest and posttest. This will be illustrated in considering the four items in the SIMS pool that concerned information transformations in geometry. The items are presented in Figure 3. The mathematics necessary to get the correct answers involves some terminology ("image", "reflection", "translation") and notation (e.g., the use of vertex letter) as well as spatial ability.

In which diagram below is the second figure the image of the first figure under a reflection (flip) in a line?

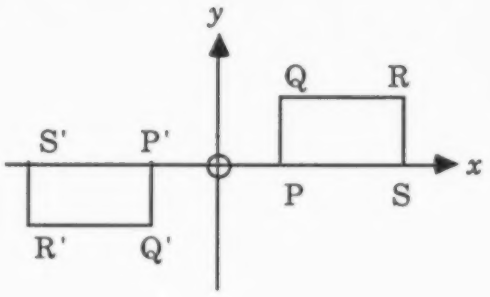
a. 

b.

c.

d.

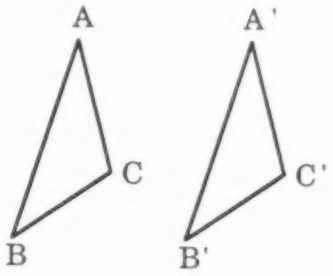
e.



PQRS is a rectangle. Its image after a transformation is the rectangle P'Q'R'S', as shown above. The transformation used could have been:

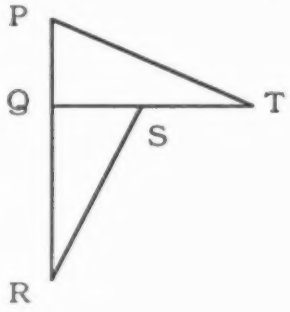
a. a rotation about the origin
b. a reflection in the *y*-axis
c. a translation parallel to the *x*-axis
d. a reflection in the *x*-axis
e. a translation parallel to the *y*-axis

A B
C D



Δ ABC and Δ A'B'C' are congruent and their corresponding sides are parallel. Δ ABC maps onto Δ A'B'C' by a

a. reflection
b. glide reflection (slide flip)
c. rotation (turn)
d. enlargement
e. translation (slide)



Δ PQT can be rotated (turned) onto Δ SQR. The center of rotation is

a. point P
b. point Q
c. point R
d. point S
e. point T

Figure 3. Four Items concerning informal transformations in geometry.

The percents of right, wrong, and omit sum to 100 percent, and so the response distribution for a given population at a given time can be plotted as a point in the equilateral triangle of a barycentric coordinate system. The corners of the triangle represent 100 percent omit, 100 percent wrong, and 100 percent correct. Each item is represented as a pair of points, corresponding to the response distribution at the beginning and at the end of the school year. The barycentric graphs for all eight countries appear in Figure 4. (Note that in Japan and British Columbia, beginning of year data were not collected for these items.) The corresponding figures, including opportunity to learn, are given in Table 1.

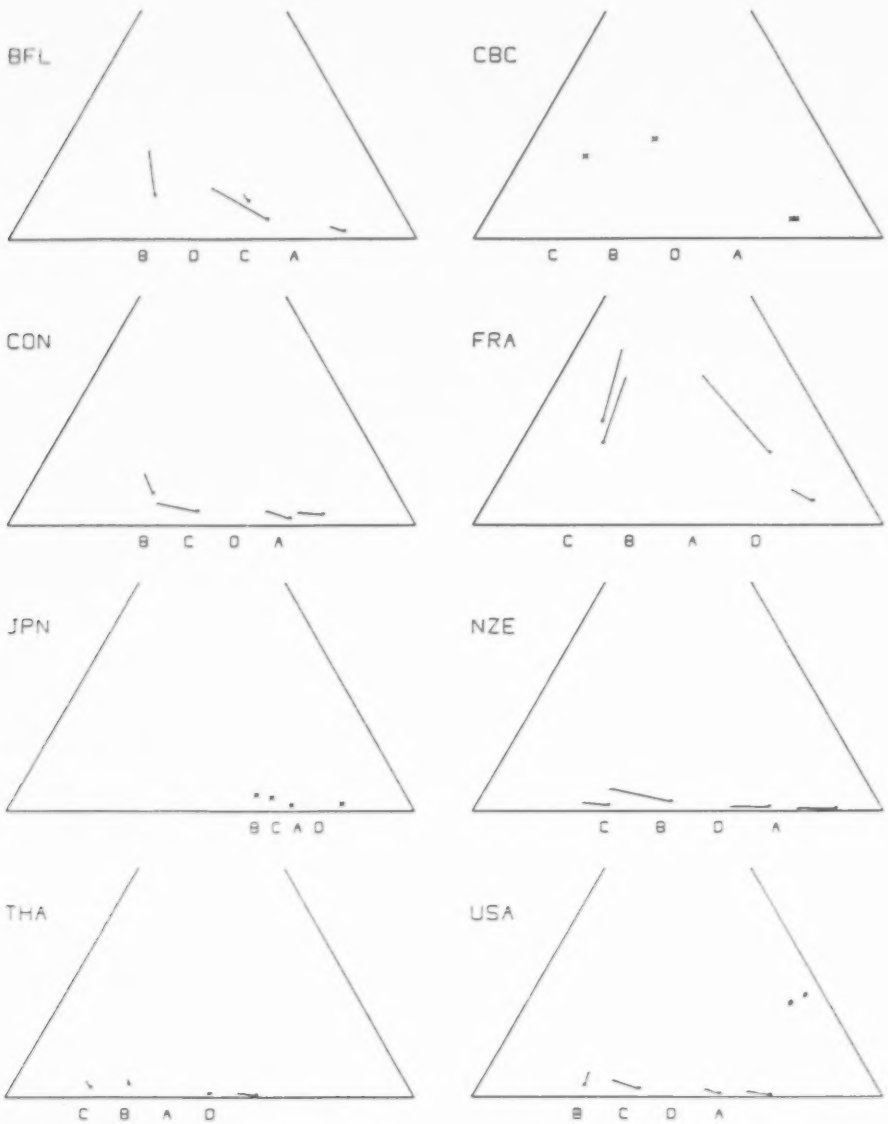


Figure 4. Beginning of the year to end of the year change in right-wrong-omit proportions for informal transformations in geometry items by country.

Note: Barycentric coordinates are used: the left corner is 100% wrong, the right corner is 100% right, and the (cut-off) top is 100% omit. In CBC and JPN, "x" shows end of year results. Otherwise, the lines show shift from befinning to end of year.

Table 1
Student Achievement and Opportunity to Learn:
Informal Transformations in Geometry

Item and Country	Student Achievement Results Pretest			Posttest			Teacher Reports of OTL		
	Right	Wrong	Omit	Right	Wrong	Omit	Previous Content	New Content	Not Taught
A (#30)									
Belgium Flemish	76	20	4	81	17	2	7	22	71
British Columbia	-	-	-	76	19	5	8	18	74
Ontario	69	28	3	76	22	3	12	29	59
France	35	23	42	62	18	20	1	85	15
Japan-	-	-	-	69	30	1	33	26	41
New Zealand	78	21	1	88	12	0	2	94	4
Thailand	50	49	1	49	50	1	6	19	75
United States	66	33	2	72	27	1	12	16	72
B (#63)									
Belgium Flemish	22	53	25	30	58	12	2	4	94
British Columbia	-	-	-	30	42	28	0	12	89
Ontario	26	59	15	31	60	9	6	22	72
France	16	42	42	20	57	23	3	21	77
Japan	-	-	-	58	37	4	9	23	69
New Zealand	30	64	6	48	50	2	1	69	30
Thailand	28	67	5	29	68	3	0	17	83
United States	25	68	7	26	71	3	2	8	90
C (#96)									
Belgium Flemish	42	44	15	60	34	5	1	22	78
British Columbia	-	-	-	16	61	23	1	12	87
Ontario	33	61	6	45	52	4	7	23	71
France	12	39	50	17	54	29	1	15	84
Japan	-	-	-	63	34	4	7	17	77
New Zealand	26	72	2	33	66	1	1	31	67
Thailand	17	78	5	20	78	3	0	36	64
United States	32	63	5	40	58	2	5	12	83
D (#158)									
Belgium Flemish	50	37	13	53	37	11	1	1	98
British Columbia	-	-	-	75	20	5	1	18	81
Ontario	61	36	4	68	31	2	5	25	70
France	72	18	10	79	14	6	2	2	95
Japan	-	-	-	81	17	2	22	23	55
New Zealand	62	37	1	72	27	1	1	73	26
Thailand	56	43	1	61	39	0	3	36	61
United States	56	42	2	60	39	1	5	13	83

Much of the story can be told by considering opportunity to learn, although the relatively high rate of correct response to some items (e.g., item A in Belgium Flemish) without apparent benefit of instructional opportunity suggests that common-sense answers were successful. In New Zealand more than other countries, the rotation and reflection items—A, B, and D—are taught and substantial growth takes place, while the translation item—C—is less often taught and less often learned. In the U.S.A., Ontario, and Thailand, there is less opportunity and less achievement. It is impossible to say whether the Japanese students acquired their high achievement or the British Columbia students acquired their mixed achievement during the year or prior to the year, since there are no pretest data.

The French and the Belgium Flemish responses are interesting because these countries have distinctive geometry curricula, involving not just informal transformations of this sort, but also formal transformational geometry. Items B and C contain the most formal terminology, and the French students show great shift in their response: they get the items wrong rather than omitting them! Items A and D involve only a little terminology, and the students show better achievement. Item A is reported by the teacher to be taught, and there is a lot of growth. In the case of Belgium Flemish, where students study vectors, only item C shows substantial growth, and that might well be explained through transfer of knowledge.

This student achievement data in geometric transformations can be compared with the teachers' opinions expressed in reaction to the proposition: "Geometry should be taught mainly through transformations (flips, turns, stretches)." The proportion of teachers agreeing or strongly agreeing was as follows:

Belgium Flemish	7%
British Columbia	3%
Ontario	8%
France	21%
Japan	26%
New Zealand	54%
Thailand	46%
U.S.A.	3%

The opinions of the New Zealand teachers especially seem to be put into practice and affect student learning, while the opinions of the Thai teachers are not in accord with the student data.

Growth in Mathematics Achievement

From the geometry analysis, we can see that achievement in mathematics and growth in achievement can be very specific: in the particular educational environment of a country, some items from a small, presumably homogeneous set are learned and others are not, and when we shift our attention to the educational environment of another country, there is a reordering of what is learned. These specificities of learning evidently depend on the specificities of opportunities to learn and on the emphasis given to different mathematical contents and perspectives. Furthermore, the psychology of the item response, or non-response, between countries and from the beginning to the end of the school year makes comparisons difficult. And this all makes us despair of our ability to aggregate the achievement results over items to form meaningful subtest scores for international comparison. Certainly a "total" score would be nonsensical.

One solution is to keep the analysis at the item level and to look over mathematical topics—and eventually over countries—for instances of high achievement and growth.

The tracking of growth will be illustrated with the results from the "core" mathematics test in the United States of America. This test consisted of 40 items stratified into 8 items from each of 5 content areas: fractions, ratio-proportion-percent, algebra, geometry, and measurement. All students were expected to take the core test at the beginning and the end of the school year. In fact, the sample size of those who did was 4399.

Because the same items were answered at each time point, the cross-tabulation can be made of right and wrong by beginning and end of year. This leads to four proportions that characterize an item's initial difficulty and its growth: the proportion of students who got the item wrong both times; the proportion of students who got the item right at first time but wrong the second time; the proportion of students who got the item wrong the first time and right the second time; and the proportion of students who got the item right both times.

These proportions sum to 1 and therefore the items can be represented in barycentric coordinates as points in a regular tetrahedron, the corners of which correspond to the hypothetical cases where 100% of the students get an item wrong at both times, 100 percent get an item right the first time but wrong the second time, etc. In order to view the configuration that the points form in the tetrahedron, the Macspin program (Donoho, Donoho, & Gasko, 1986) was used. This program runs on the Apple Macintosh and allows a three-dimensional configuration to be viewed as it rotates around any axis. As soon as motion begins, the eye forms a good picture of the configuration. The static, two-dimensional projections given in Figures 5a and 5b are snapshots taken from several views. Although there are three degrees of freedom in the item statistics, the points closely follow a two-degree-of-freedom surface. The program was used to focus on that surface, in Figures 5c and 5d, and then to label the points, in Figure 6.

One major finding is that growth is small. The reason is certainly not that there is no room for growth, since most of the beginning-of-year results are in the lower or moderate category. There are a few items with spectacular gains, but this provides little comfort when we look at the content of these items. The item providing the largest gain has the following stem: "(-2) x (-3) is equal to...." That is, students do not know the multiplication rule for negative numbers at the beginning of the year, but they do seem successfully to learn it. The second highest gainer is: "if X = -3, the value of -3X is...." which is the same rule, with a little notation.

When the points are tagged with content categories, we see again the fact that there is great inhomogeneity of achievement within what was considered to be homogeneous content units.

When the points are divided according to high and low opportunity to learn, the effect of instruction is evident.

Conclusions

The major finding of the SIMS analysis and survey of mathematics achievement and growth at critical juncture between elementary and secondary education is that not very much mathematical achievement is taking place. We do see some rather direct connections between curriculum and learning, and so perhaps the conclusion should be that the objectives of the mathematics curriculum are too limited: if more content were introduced, it seems likely that more mathematics will be learned. Furthermore, from analyses in Burstein (1989), we know that the attitudes of the students—shared to a great extent by their teachers, and not undergoing very much shift during this year of school—are that mathematical formulas and rules and the calculation of answers are what is important. Perhaps if mathematics were cast in a more creative and interesting light, students would like it better and would be motivated to learn more.

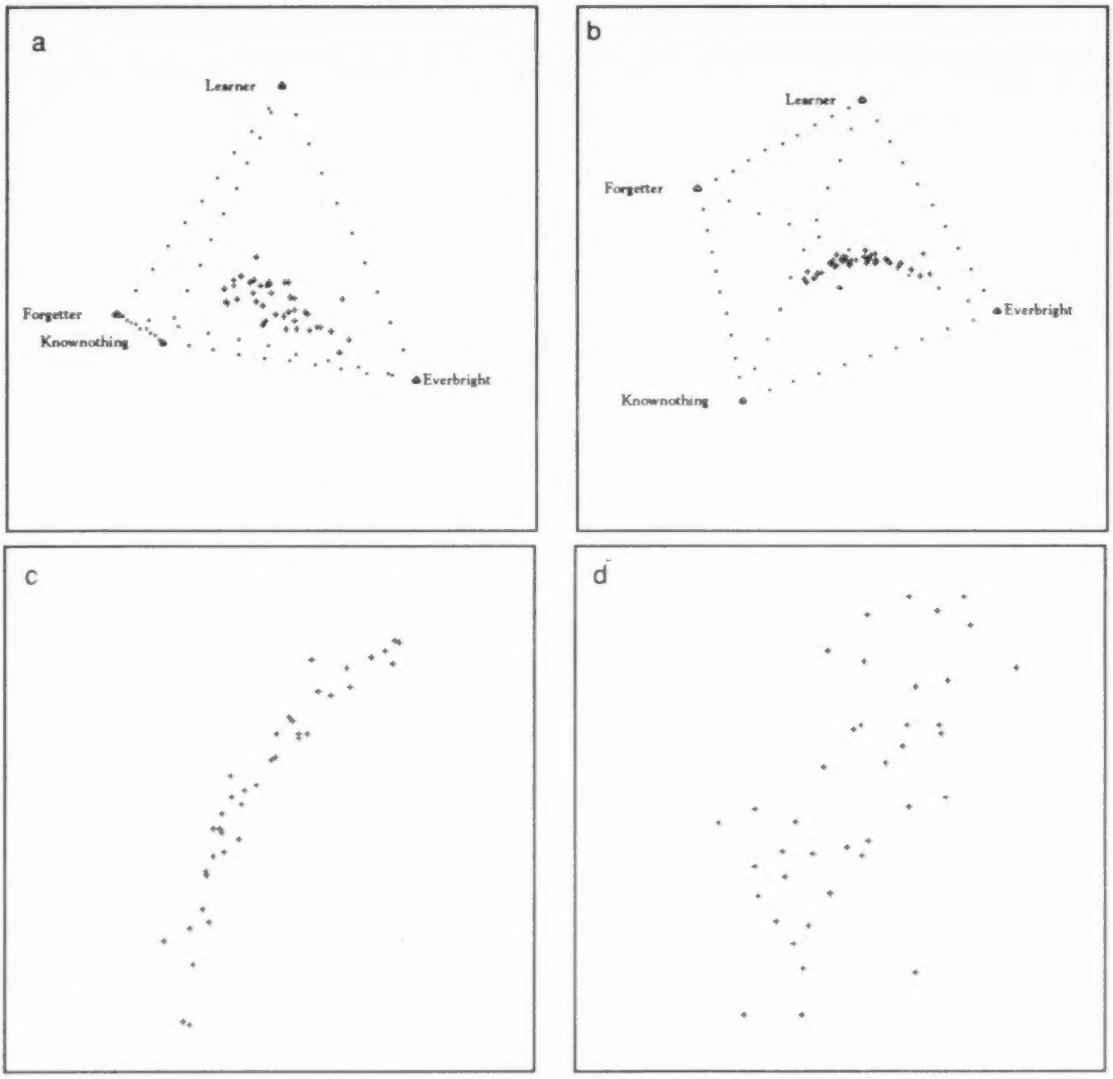


Figure 5. Not knowing, knowing, learning, and forgetting mathematics items.

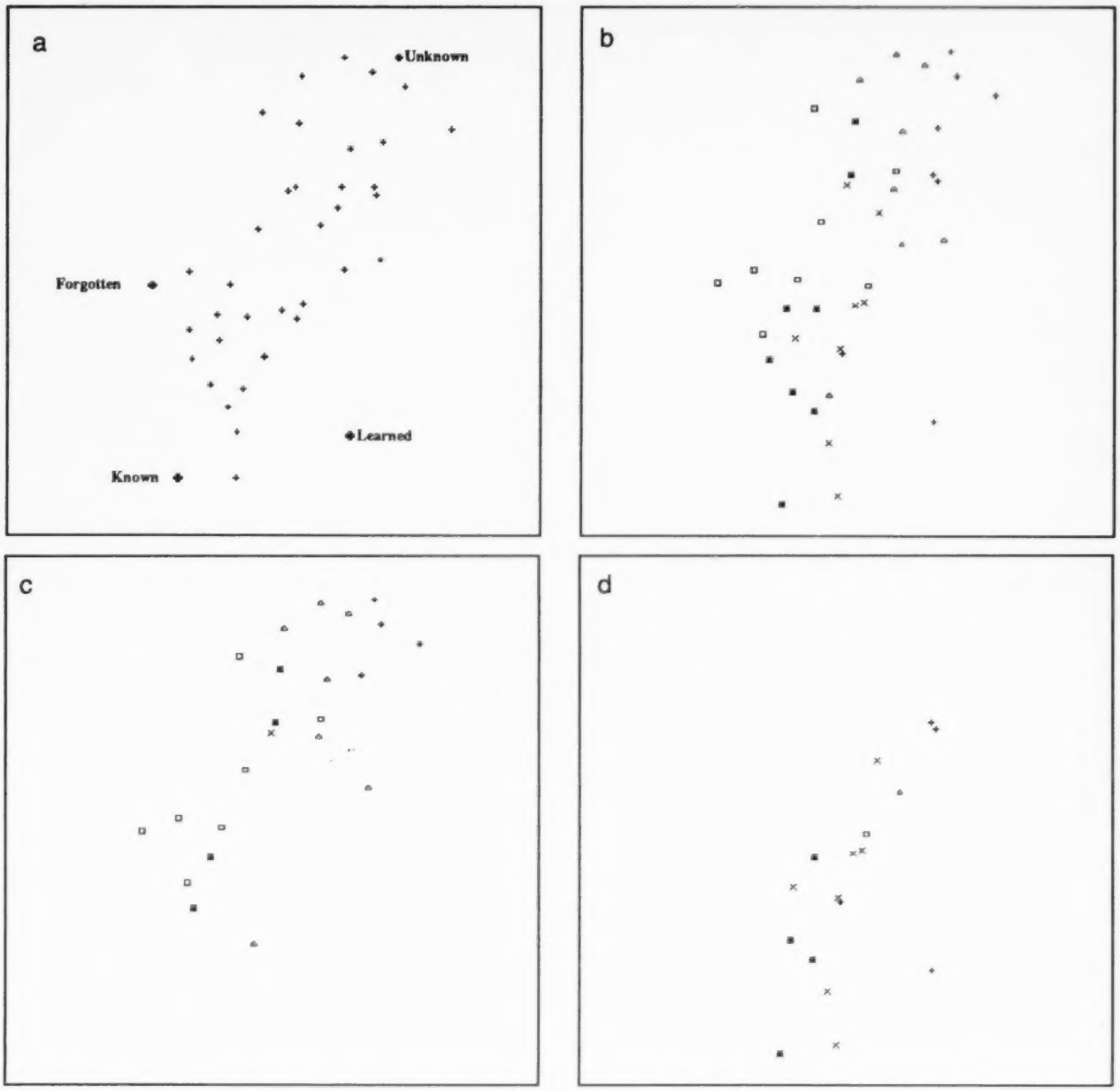


Figure 6. Item not knowing, knowing, learning, and forgetting by content and opportunity to learn.

Note:
a. "Unknown" is an item rarely learned. "Known" is an item usuallly known at the beginning of the year. "Learned" is an item often learned during the year. "Forgotten" is an item often forgotten during the year..
b. By content of item: ❑ is measurement; Δ is geometry; x is ratio-proportion-percent; ● is fractions.
c. Less than 75% opportunity to learn this year.
d. More than 75% opportunity to learn this year..

References

Burstein, L. (ed.). *Second international study of achievement in mathematics: Student growth and classroom processes in lower secondary school.* To be published in 1989 by Pergamon Press.

Donoho, A. W., Donoho, D. L., & Gasko, M. *Macspin graphical data analysis software*. D2 Software, Inc. P.O. Box 9546, Austin Texas 78766 USA.

Husén, T. (1967) *International study of achievement in mathematics: A comparison of twelve countries*. Volumes I and II. New York: John Wiley & Sons.

Travers, K., & Westbury, I. *Second international study of achievement in mathematics: Analysis of the international mathematics curriculum* To be published in 1989 by Pergamon Press.

Part 3

National Initiatives in Evaluation

THE CURRICULUM OF THE *SCUOLA MEDIA* SINCE 1979

Raimondo Bolletta

In this paper I will be presenting the results of a research project called VAMIO (*Verifica Abilità Matematiche Istruzione dell'Obbligo* - Verification of Mathematical Ability in Compulsory Schooling) which was developed through a course of study for a doctorate in experimental educational research. The project was conducted with funds from the European Center of Education in Frascati.

With the third and final year of the *Scuola Media* (lower secondary school), 13-14 year olds finish the 8-year program of compulsory schooling in Italy. The school system in Italy is centralized and there is a national syllabus for each age group up to 14 years of age. There is also a national syllabus for each kind of upper secondary school, but our assessment system is rather informal and not centrally controlled. We do not have any kind of examination board and all the examinations and all types of evaluation throughout schooling are directly administered by classroom teachers. The only form of external evaluation occurs at the end of the upper secondary school for the final Diploma (*Maturità*). A commission of external teachers appointed by the Ministry of Education assesses the student's achievement on the basis of two written essays and on the outcome of an interview covering 4 disciplinary subjects: two chosen by the commission, and two by the individual student.

There are two main consequences of this situation. The first one is positive, in that there is a lot of freedom, and it is possible to introduce any kind of innovative teaching experiment that we want (more or less). On the other hand, it is very difficult for the school system to have documented knowledge of what is really happening in different parts of the country, and to use the same kind of measure and the same standards nationally for the outcomes of the school system.

In 1979 the Ministry of Education carried out a reform of the *Scuola Media* which changed the syllabuses in all subjects. The new programs in mathematics were particularly innovative. These programs were prepared by a large commission in which many promoters of innovation, both secondary school teachers and university professors, were represented.

It is difficult to summarize in a few words all the rich and interesting aspects of these programs. I shall mention only a few. Mathematics and experimental science are taught by the same teacher. Programs are not prescriptive but they suggest some general themes and subthemes. The classroom teacher is given direct responsibility for the choice of specific topics in each area and for the organization and scheduling of classroom activity. Topics such as probability and statistics, logic and introduction of geometry by isometric and non-isometric transformations are the main innovations (from the point of view of contents), whereas from a methodological point of view, particular attention is paid to interdisciplinarity, to the applications of mathematics to reality, and to some simplifications of algebraic rules of calculation. Set theory is recommended only as a language among others.

Three years later, it was deemed necessary to change the final examinations of the *Scuola Media* in order to make them correspond to the new contents and methodology. This gave rise to a large debate on the best ways to assess changes in students' performance and, more generally, on the problem of the effectiveness of the new programs.

Wide-spread discontent was perceptible: from teachers of the *Scuola Media* because the syllabus was too ambitious and too vast, and from upper secondary school teachers because levels of achievement of students were decreasing. There was a general agreement on the fact that it was very difficult to implement revised programs in the *Scuola Media* if both elementary school (last reform in 1952) and upper secondary school programs remained antiquated.

The idea of the VAMIO survey sprang-up in this context, and considered these kinds of problems. The principal aim was to produce a standardized test to help evaluate levels of achievement of single classes or individual students at the end of the *Scuola Media* and to diagnose the real preparation of students and the eventual need of remedial work at the beginning of upper secondary school. In order to reach this aim, it was necessary to investigate the effectiveness of the programs more deeply, and to know something more about the actual implementation of the mathematics curriculum.

The problem was to find a simple way to collect data on the actual interpretation and implementation of the official programs. For this purpose, we based ourselves upon the methodology of the IEA surveys, in particular, upon the preliminary studies of intended and implemented curricula.

Is it possible to have reliable indications about actual activities inside classes directly from teachers? Are they good judges and impartial observers of the class situation? How should one define, elaborate and use the variable "Opportunity to Learn"? How should

one describe in a clear and understandable way the content of programs? Is it possible to have instruments for measuring the amount of innovation promoted and implemented by the program?

Sample

We interviewed a national sample of 1300 teachers by a questionnaire on the "actual" program implemented in the classroom and, one year later, studied the achievement of an independent sample of 2800 students by means of a multiple-choice test.

It seemed that the crucial variable was the teacher. We interviewed a representative sample of mathematics teachers of the *Scuola Media* by questionnaire. Each teacher was asked about three clusters of variables: the first one refers to teacher characteristics (sex, age, type of degree, place of residence, textbook used, general teaching attitudes); the second one concerns the program actually developed in the classroom; and the third one is the "opportunity to learn" which referred to a set of items.

The scholastic program was described through a list of contents, about 150 topics, and for each one of them the teacher had to tell its relevance in terms of time spent in class to develop it, in which grade it was normally developed, and the level of difficulty for students to learn it. The relevance variable was expressed by a six-value scale (0 = the content was not taught; 1 = brief comments in one or two lessons; 2 = general but synthetic treatment; 3 = thorough treatment in 5-8 lessons, even in different years; 4 = systematic and repeated treatment; 5 = the content was developed with particular care in 20-30 lessons during the three-year course of the *Scuola Media*). As a control of these indications, for a set of about 140 items, we have the values of the "opportunity to learn" variable which is defined as the predicted percentage of students able to answer the items correctly.

Teachers responded to this survey positively. 89 percent of the schools invited to participate accepted and 91 percent of the teachers inside the accepting schools answered the questionnaire. Controls of coherence among different variables show that we had a good quality of answers. In particular, it does not seem that teachers gave a biased or optimistic image of the real activity in the classroom.

Keeping in mind only distributions and modal values of the relevance variable, it is possible to have an interesting map of the syllabus which split the contents into three clusters: the first containing topics whose modal value of importance is 4, the second one containing topics whose modal value is 0 or 1, and the third containing the remainder of the list. Looking at these three parts of the list, it seems that actual syllabuses are considered too vast and each teacher decides what part of the program he or she should develop. Although a large majority of teachers are in agreement with a core program which contains the most traditional topics, they eliminate many topics that are too innovative or too difficult for most students (excluded program). The remaining topics are developed optionally, only by small numbers of teachers.

Table 1
Topics of Syllabus Most Often Covered (Core Syllabus)

Topic	Frequencies 0	1	2	3	4	5	
GEOMETRY- (THE FIRST REPRESENTATION OF THE PHYSICAL WORLD)							[A]
01 Study of plane figures arising from models of: nature	2	19	18	19	33	10	
03 Drawing plane figures	1	14	19	17	41	7	
04 Nomenclature relative to polygons	1	25	25	12	33	4	
06 Calculation of perimeters and areas of quadrilaterals.	2	1	3	17	57	20	
10 Figures with equal areas	1	4	14	22	47	13	
12 Study of regular polygons	1	7	24	28	32	8	
13 Theorem of Pythagoras	0	1	6	21	50	22	
14 Application of Pythagoras' Theorem in the solution of geometric problems	0	0	1	10	54	36	
15 Use of straight edge, square and compass in geometric constructions	12	27	17	10	29	6	
24 Study of solid figures arising from models of nature	1	17	21	16	39	5	
25 Regular polyhedra	5	21	23	15	33	4	
28 Cube	1	6	26	23	40	3	
30 Parallelepiped	0	5	25	23	44	4	
31 Prism	1	5	24	24	44	3	
32 Pyramid	1	5	21	26	45	4	
33 Cylinder	1	4	23	24	44	4	
34 Cone	2	4	23	24	44	4	
37 Composite solids	7	11	20	19	38	5	

Topic	Frequencies 0	1	2	3	4	5	
NUMERICAL SETS							[B]
01 Set of natural numbers	1	14	22	19	35	10	
03 Decimal metric system	1	13	26	23	33	5	
05 Operations with signed numbers	1	1	7	24	56	11	
06 Comparison of signed numbers	1	15	24	20	34	6	
07 Graphical representation of signed numbers	0	16	25	18	35	5	
08 Fraction as an operator	2	4	10	25	45	14	
09 Equivalent fractions	1	8	20	26	39	7	
10 Concept of ratio	1	5	17	24	44	10	
12 Expressions with rational numbers	1	3	15	23	46	11	
13 Proportions	0	1	7	29	53	10	
14 Solving for an unknown in a proportion	0	5	17	27	45	6	
15 Application of proportions in the solution of problems	1	2	10	23	54	10	
22 Direct and inverse numerical operations	1	7	17	21	43	11	
23 Properties of numerical operations	1	8	22	27	35	7	
24 Raising to a power	0	4	17	34	39	6	
27 Common multiples and common divisors of several numbers	1	5	23	33	34	5	
28 Prime factorization.	0	4	21	36	35	4	
29 Rules for the calculation of the GCD and LCM	0	4	20	36	35	4	
30 Exercises in exact and approximate calculation	8	18	24	17	28	5	
32 Effective use of numerical tables	2	16	26	20	33	4	
MATHEMATICS OF CERTAINTY AND MATHEMATICS OF THE PROBABLE							[C]
PROBLEMS AND EQUATIONS							[D]
01 Recognition of significant information and variables in a word problem	3	5	10	14	44	24	
03 Setting-up of arithmetic expressions for the solution of a word problem	5	15	26	21	27	7	
05 Reading, writing, use and manipulation of simple formulas	3	4	11	16	49	18	
06 First-degree equations	1	1	8	29	49	11	
COORDINATE GEOMETRY							[E]
01 Coordinate geometry in concrete situations	7	19	22	19	27	7	
04 Coordinates of a point in the plane	2	10	25	27	30	5	
06 Cartesian plane representation of mathematical laws describing real phenomena	5	6	23	26	35	5	
08 Cartesian representation of direct proportionality	3	4	20	29	39	6	
09 Cartesian representation of inverse proportionality	3	4	20	29	39	5	
GEOMETRIC TRANSFORMATIONS							[F]
CORRESPONDENCES AND STRUCTURAL ANALOGIES							[G]
03 Concept of function	7	9	25	25	29	6	

TABLE 2
Topics of the Syllabus Rarely Covered (Excluded Syllabus)

Topic	Frequencies 0	1	2	3	4	5	
GEOMETRY THE FIRST REPRESENTATION OF THE PHYSICAL WORLD							[A]
02 Construction of triangles with material	10	43	30	12	5	1	
05 Using Venn diagrams with sets of polygons	21	37	22	8	10	1	
07 Symmetry of the square	17	39	27	9	8	1	
08 Symmetry of quadrilaterals	21	34	28	9	7	1	
09 Axes of symmetry of triangles	18	35	29	9	8	1	
11 Convex and concave polygons	11	57	20	5	6	1	
16 The problem of calculating *pi*	10	41	26	9	12	1	
17 The problem of squaring the circle	50	28	16	4	3	0	
21 Relative positions of two lines in space	5	42	34	11	7	1	
22 Dihedral angles	7	46	31	8	7	1	
23 Angles in a solid	41	31	18	5	5	1	
26 Axes of symmetry of regular polyhedra	44	27	15	7	7	1	
27 Euler's formula for a polyhedron	59	23	10	3	4	1	
29 Plane sections of the cube	33	24	18	10	14	1	
35 Spheres	26	16	20	15	21	2	
36 Plane sections of the cone and the cylinder	34	23	18	10	14	1	
NUMERICAL SETS							[B]
02 Ancient number systems	8	60	24	4	4	0	
04 Arithmetic of odd and even numbers	24	35	25	7	8	2	
19 Base 2	21	24	34	15	6	0	
20 Bases other than 10	27	27	30	11	5	0	
21 Order of magnitude	9	34	30	12	15	1	
31 Successive approximations as an approach to real numbers	32	26	23	10	8	2	
33 Use of calculators	39	27	16	7	9	2	
MATHEMATICS OF CERTAINTY AND MATHEMATICS OF THE PROBABLE							[C]
02 Logical connectives	43	18	21	10	7	1	
03 Circuits and switches	49	16	20	11	4	1	
04 Logical operations and set operations	36	14	23	15	10	1	
09 Density maps	29	21	24	14	11	2	
10 Absolute frequency	30	27	26	12	6	1	
11 Relative frequency	30	25	27	12	6	1	
14 Surveys	29	18	25	15	11	2	
15 Phases in a statistical study	35	21	24	12	8	1	
16 Discrete and continuous variables	70	13	11	4	3	0	
17 Time series	82	7	6	2	3	0	
18 Various types of tables	41	18	19	9	12	2	
19 Mode	45	2	19	7	5	1	
20 Median	44	23	20	8	5	1	
21 Weighted arithmetic mean	58	17	15	6	4	0	
22 Experimental laws and interpolation	70	10	11	5	4	0	
23 Sampling	58	20	13	4	4	1	
24 Statistics and probability	29	17	27	15	10	2	
25 Simple geometric mean	64	17	12	4	3	0	

Topic	Frequencies						
	0	1	2	3	4	5	
26 Properties of the arithmetic mean	68	14	11	4	2	0	
27 Dispersion	83	9	4	2	1	0	
28 Index numbers	84	8	5	2	1	0	
29 Gauss' curve	57	24	13	3	2	0	
30 Extrapolation	81	11	5	2	1	0	
31 Graphic representation of polar coordinates	83	7	5	3	2	0	
32 Properties of the median	80	11	5	2	2	0	
33 Correlation	86	7	5	1	1	0	
34 Tables of random numbers	90	5	3	1	1	0	
35 Structure of populations by age grouping	78	12	6	2	2	0	
36 Rate of growth of population	72	16	7	3	2	0	
37 Characteristics of census taking	70	19	7	2	2	0	
38 Frequency	53	22	14	6	4	0	
PROBLEMS AND EQUATIONS							[D]
02 Flowcharts	26	19	21	13	17	4	
07 First degree inequalities	48	14	17	11	9	1	
COORDINATE GEOMETRY							[E]
02 Reading topographical and geographical maps	22	30	22	13	11	2	
05 Representation of polygons on graph paper	14	11	27	23	21	4	
07 Graphical representation of exponential growth	45	17	19	11	8	1	
10 Cartesian graph of $y = x^2$	23	11	23	18	21	3	
14 Condition of perpendicularity of two lines	36	13	23	13	13	2	
15 Graph of an inequality	79	6	8	4	3	1	
16 Applications to problems of linear programming	82	6	6	3	3	1	
17 Analytical study of conic sections	84	6	5	3	3	0	
GEOMETRIC TRANSFORMATIONS							[F]
02 Use of protractor	3	35	29	15	16	2	
03 Construction of an angle bisector	11	41	29	9	9	1	
09 Set of isometries and compositions of isometries	53	12	18	10	7	1	
10 Dilations	45	16	21	10	8	1	
15 Observation of shadows in the plane	58	14	16	7	5	0	
16 Properties of affine transformations	74	10	10	4	3	0	
17 Equations of affine transformations	86	5	5	3	2	0	
18 Equations of similarity transformations	79	6	8	4	2	0	
19 Equations of symmetry with respect to the Cartesian axes or the origin	73	9	10	5	3	0	
20 Drawing in perspective	84	8	4	3	1	0	
21 Deformed images	85	8	4	2	1	0	
CORRESPONDENCES AND STRUCTURAL ANALOGIES							[G]
04 Search and discovery of structural analogies	55	9	11	9	14	2	

Table 3
Topics in the Syllabus of Intermediate Emphasis (Optional Program)

Topic	Frequencies					
	0	1	2	3	4	5
GEOMETRY THE FIRST REPRESENTATION OF THE PHYSICAL WORLD [A]						
18 Inscribed and circumscribed polygons	1	12	32	30	23	3
19 Lines tangent to a circle	4	36	37	13	9	1
20 Inscribed angles and central angles	4	26	43	16	10	1
NUMERICAL SETS [B]						
11 Percentages	1	8	28	31	27	4
16 Representation of rationals on the number line	5	27	28	15	23	2
17 Decimal form of rational numbers	2	14	35	26	21	2
18 Terminating decimals and repeating decimals	2	10	37	30	20	1
25 Rules for extracting a square root	9	15	29	26	18	2
26 Methods of approximation of a square root	4	15	36	24	19	3
MATHEMATICS OF CERTAINTY AND MATHEMATICS OF THE PROBABLE [C]						
01 True-false statements and probable statements	15	25	30	14	13	3
05 Statistical observation	17	16	29	21	13	3
06 Pie charts	8	21	32	22	16	2
07 Pictograms	9	26	31	18	15	2
08 Histograms	7	21	30	23	17	3
12 Percentages	7	11	30	28	21	3
13 Simple arithmetic mean	11	29	33	16	10	1
PROBLEMS AND EQUATIONS [D]						
COORDINATE GEOMETRY [E]						
03 Coordinates of a point on a line	3	16	28	23	26	4
11 Equation of a line through the origin	10	11	29	21	26	4
12 Equation of a line parallel to an axis	17	13	30	18	20	3
13 General equation of a line	16	11	30	18	23	2
GEOMETRIC TRANSFORMATIONS [F]						
01 Measure of angles	0	9	31	33	24	3
04 Sum of internal angles and external angles of a triangle	1	19	41	23	15	2
05 Rigid motions in the plane	24	17	27	18	12	2
06 Translations	25	16	28	17	12	2
07 Rotations	20	16	29	18	15	2
08 Symmetries	16	18	31	19	15	2
11 Similarity	10	8	26	32	23	1
12 Properties of similar figures	10	8	26	34	21	1
13 Relationship of areas of similar figures	12	9	30	29	19	1
14 Scale drawing	11	16	30	24	17	2
CORRESPONDENCES AND STRUCTURAL ANALOGIES [G]						
01 Concept of relation	16	17	25	18	20	5
02 Concept of correspondence	13	15	28	20	21	5

This situation is supported by the results of students: 60-70 percent of the students answered correctly items which referred to topics included in the core program, whereas only 25-30 percent of students succeeded on items related to topics eliminated from the official syllabus. Responses during the administration of the trial test of 450 students gave a further confirmation of this: for each item they specified if they had already studied this particular topic. We found the same kind of agreement with the mean value of "opportunity to learn" (OTL) variable.

With Freudenthal's criticisms in mind (Freudenthal, 1975), particular attention has been reserved for the analysis of OTL. This variable was presented as the predicted percentage of students able to correctly answer an item and, in this sense, as a measure of the teacher's ability to predict the achievement of students. But at the same time, as a curriculum indicator, assuming that the time spent in the classroom is positively correlated with student achievement. To check the first interpretation of the variable —measure of ability to predict the achievement of students— for each item of the test, we studied the contingency table between the OTL expressed by the 140 teachers, and the mean score obtained on the item by the entire class of each particular teacher, and we calculated the value of chi-square. But we found that the value of significance of chi-square depended on the item, so we had to find another way to categorize them. Items referring to topics in the core program show a dependence between the achievement of the class and the OTL expressed by the teacher, whereas items referring to excluded or optional topics or items with hidden difficulties or poorly formulated items present the achievement of the class independent of the teacher prediction.

I must also mention another statistical aspect of the OTL. For each item, there is a strong stability of mean for the OTL in both independent samples of teachers: the 1300 teachers interviewed in 1985 and the 140 teachers of the classes tested in 1986.

In order to classify the educational options indicated by the list, we tried to reduce the number of variables in play. This analysis of the information relative to the curriculum actually covered consists in a factor analysis of the relevance variable.

The Achievement of Students

The test used in the VAMIO research did not try to propose criteria for evaluating the quality of the innovation actually realized. In fact, as much as possible, it tried to avoid proposing a particular interpretation of the syllabus.

A qualitative analysis of each item and of its statistical characteristics allowed us to discover different levels of preparation of students in different parts of the syllabus and this has been a check on the information collected through the teacher questionnaire. The results amply confirm what had already emerged from the analysis of the syllabus. But the analysis of errors and the factorial analysis of the test also suggested some didactical problems: for example, it seems that the ability to read and interpret a statistical diagram is independent of the ability to work in the Cartesian plane.

Different parts of the program are not well integrated. Due to the fact that some parts are considered as optional, we found that if one part is well-developed by a particular teacher, another is less so and vice-versa. For example, two items which refer to numerical ability (the first concerned with the structural properties of numerical sets while the second referred to the approximate result of a multiplication) correlate in opposite ways with the same factor. It seems that the diverse nature of the items is amplified by the different didactic options which are compatible in the same program.

Test results also demonstrated the existence of significant differences among students from different geographical regions. Students from the more industrialized and wealthier North scored higher than their counterparts in the South. This fact, which was already evident in previous IEA studies (Laeng, 1977; Visalberghi, 1977), suggested a further analysis of data regarding the implemented curriculum. Comparing the mean value of the ratings of the factors, those related to the most innovative topics rated higher in the North, while those related to traditional topics rated higher in the South.

On the basis of this experience, I think that:

1. it is possible to survey the implementation of a centralized program using cheap and quick instruments for collecting data directly from teachers;
2. in our particular situation in Italy, we must gather and analyze more information about the achievement of large samples of students;
3. it is not possible to introduce innovation simply by writing good syllabuses. We need, to consider the entire educational process, in order to control trends of interpretation, attitudes of teachers, and the achievement of students.

4. the actual national program constitutes a conceptually advanced proposal not yet fully developed or actuated. More energy (time and money) must be invested in in-service teacher training, development of educational materials, and research.

References

Bolletta, R. (1987) *Il rendimento in matematica alla fine della scuola dell'obbligo: costruzione e validazione di un test oggettivo per l'accertamento finale e la diagnosi di ingresso degli studi successivi*. Unpublished doctoral thesis, Universitá La Sapienza di Roma.

Bolletta, R. (1988). *La preparazione matematica in Italia alla fine della scuola media*. Frascati, Centro Europeo dell'Educazione.

Freudenthal, H. (1975). Pupils' Achievements Internationally Compared - The IEA. *Educational Studies in Mathematics*, 2, 127-187.

Laeng, M. (1977). I risultati IEA e le prospettive di innovazione nell'insegnamento delle scienze e della matematica. In Misurazione del rendimento scolastico: indagine IEA e situazione italiana. *Quaderno degli Annali della Pubblica Istruzione*, 5, Roma.

Visalberghi, A. (1977). Rapporto generale sui risultati IEA in Italia e sulle ricerche connesse. Valutazione complessiva dei risultati. In Misurazione del rendimento scolastico: indagine IEA e situazione italiana. *Quaderno degli Annali della Pubblica Istruzione*, 5, Roma.

THE 1987 APU SURVEYS: SOME PRELIMINARY RESULTS

Derek Foxman · Graham Ruddock

The 6th national mathematics monitoring surveys of 11 - and 15-year-olds in England, Wales, and Northern Ireland were carried out in 1987 by the National Foundation for Educational Research in England and Wales, on behalf of the Assessment of Performance Unit at the Department of Education and Science (DES) in Britain. As in the previous phase of annual surveys from 1978 to 1982, a light sampling technique was used in which each pupil involved took only a fraction of the total assessments used. Modes of assessment relating to new technology and small group problem solving were included for the first time as well as the modes used in previous surveys. Some initial analyses of the age 11 survey data have been carried out which show a similar picture to 1982 in the pattern of results with some shifts in detail.

Surveys of the mathematical performance of pupils in the 11-and 15- year old age groups in the schools of England, Wales and Northern Ireland are being carried out by the National Foundation for Educational Research in England and Wales (NFER). The NFER is an independent research body funded mainly by the Local Education Authorities in England and Wales and by outside sponsors. It undertakes research and development projects on issues of current interest in all sectors of the public education system.

The monitoring surveys are conducted on behalf of The Assessment of Performance Unit (APU) at the DES and are funded by the DES, The Welsh Office Education Department, and the Department of Education in Northern Ireland. In the APU's monitoring programme the NFER has also undertaken surveys in English Language and First Foreign Language, work in science is based at the University of Leeds and Kings College, London University, while Goldsmiths College, London University, is responsible for Design and Technology.

The research teams' work is guided by steering groups consisting of teachers, education authority curriculum advisers, educational researchers and members of Her Majesty's Inspectorate (HMI). The work of steering groups is supervised by a management team of the APU consisting of a small number of administrators and HMI.

The purpose of the surveys is to provide a national picture of the performance of pupils in the age groups concerned and, over a series of surveys, to monitor changes in performance.

Six mathematics surveys of each age group, 11 and 15, have been mounted. Both groups were surveyed annually from 1978 to 1982, and a further survey was carried out in 1987. Surveys of 11 year olds are carried out in May, with those for the older pupils taking place in November. Age 11 represents the last year of primary schooling, while the 15 year olds, in November, are at the beginning of their last year of compulsory schooling.

The Mathematics Assessment Framework

The assessment framework on which the mathematics surveys are based can be seen as having three dimensions:

Content. The mathematical content is covered by five main categories: number, measures, geometry, algebra, probability and statistics. Each of these categories is further divided into over a dozen sub-categories in toto for detailed monitoring purposes. This division into sub-categories differs for the two age groups; number is, for example, represented by a greater number of divisions at age 11, than at age 15 while for algebra the reverse is true.

Context. The contexts in which the mathematics is placed includes everyday life, other school subjects, such as geography, and that of mathematics itself.

Learning Outcome. Three broad forms of learning outcome are assessed: the understanding of concepts and performance of routine skills; using problem solving strategies and attitudes to mathematics.

Modes of Assessment: The 1987 Surveys

In the 1987 surveys the assessment modes already developed in the 1978 to 1982 period were again used together with specially developed new modes which reflected trends in the mathematics curriculum since 1982.

The modes of assessment used from 1978 to 1982 were:

- **Written tests of concepts and skills.** Each test comprised around 50 short response items. Only a few of these were multiple-choice questions. Normally

three sub-categories of content were equally represented in each test. Up to 30 such tests were used in a survey.

• **Written tests of problem solving strategies** (Problems & Patterns): five or six problems were presented in each test. A graded set of questions was presented for each. Between 8 and 10 tests were used in each survey.

• **Written attitude questionnaires:** Sections on attitudes to mathematics in general and to particular topics were presented. The scales related to the enjoyment, usefulness and utility of mathematics and mathematical topics. A free response section was also included.

Concepts and skills, problem solving strategies, and attitudes were also assessed in practical tests in an oral mode given in a 1-to-1 interview situation by experienced teachers of the age group, recruited and trained specially for a survey. The training aimed to produce a high degree of standardisation of presentation, but with some flexibility allowed in a friendly, but searching atmosphere. Assessors worked from a "script" containing all the questions to be put to pupils and directions on the manner in which the materials to be used should be presented. Prompts or hints were given if prescribed by the script. Flexibility was provided by the freedom given to assessors to ask for clarification of a response in a neutral way whether scripted or not. For example, pupils could be asked how they obtained an answer whether the response is correct or incorrect. Each pupil was given about three topics from up to 15 used in a survey. Assessors recorded in as much detail as was practicable what they and the pupils said and did.

The overall balance of assessments in 1987 as compared with 1982 shifted towards more problem solving and mathematics in context and the greater use of new technology, calculators, and microcomputers.

A larger number of Problems and Patterns and fewer Concepts and Skills tests were used than in 1982. Within each assessment mode the role of the calculator was increased. A new number sub-category, Calculator Skills, was introduced into the Concepts and Skills assessments, the existing sub-categories remaining non-calculator based. In 1978 to 1982 calculators had not been allowed in any of the Problems & Patterns tests whereas in 1987 they were allowed in about one-third of them. In addition, more of the topics in the Practical Tests were calculator based. All of the collections of items were reviewed, revised and updated.

New assessment modes were also introduced for 1987:

• **Written Theme Tests:** In the theme tests a unifying context, such as the weather, or planning a trip, was provided to produce a meaningful setting for the tasks. These consisted of short response items together with a final task requiring integration of a range of information and previous answers. Calculators were available.

• **Small Group Problem Solving Tasks:** Groups of three pupils of the same sex and similar attainment level worked together on problem solving tasks with an assessor recording the activities (Foxman, in press).

• **Mathematics with the Micro:** Individual pupils undertook problem-solving activities on a BBC B or

Table 1
Structure of the 1987 Primary Mathematics Survey

	Sample used	No. of Pupils
Concepts and Skills	Whole sample	10,000
Written tests: Problems and Patterns Theme tests Calculator Skills test	Sub-sample 1	4,800
Attitude Questionnaire	Sub-sample 2[a]	1,200
Practical tests	Sub-sample 3[a]	1,200
Small Group Problem Solving[b]	Separate sample (270 groups of 3)	800
Maths with the Micro[b]	Non-random separate sample	250

[a] Sub-samples 2 and 3 overlap
[b] These pupils also took a specially constructed Concepts and Skills test.

RML Nimbus microcomputer. An assessor recorded the activity on a 1-to-1 basis. This was a probe, using a non-random sample, to illustrate the use of a microcomputer for such assessments.

Survey design in 1987

Stratified cluster sampling is used for all random samples. About 2 percent of the target population is sampled in England and about 6 percent in Wales and Northern Ireland. The survey designs for the two age groups were similar; that for the younger pupils is shown in Table 1 by way of example.

All pupils took a Concepts and Skills test, and most, in addition, took a second assessment; a few 11-year-olds took three. Since an overall picture of performance is required, each pupil took only a small proportion of the items in use, and made his or her contribution to the overall picture by so doing. A large and representative pool of items can thus be widely sampled. All assessments are taken anonymously. Schools in the sample are asked to complete questionnaires about some aspect of their curriculum and their staffing resources. Other information is requested about the size of classes and their methods of organising them, especially mathematics classes. The amount of time spent on mathematics in school and doing homework is also asked for. Further information about schools' location, size and pupil-teacher ratio is obtained from the DES.

Reporting

The results are reported in a number of ways. For monitoring purposes reporting has been by sub-category for the Concepts and Skills tests and against some pupils and school variables. The reporting by topic or individual item, however, has been more valuable for teaching purposes. A multi-level modelling programme is being used to analyse the 1987 data by background variables (Goldstein, 1986; Hutchison and Schagen, 1987).

The form of reporting for the 1987 surveys has not yet been finalised. Previous reports have included individual reports on each of the first three annual surveys at each age level (Foxman et al. 1980, 1981, 1982), and also a Review Report covering the findings of all annual surveys (Foxman et al. 1985). These reports cover every aspect of the survey and are not written for particular audiences. As the APU's programme has progressed, the emphasis in the research has shifted from an overriding interest in monitoring change to obtaining and disseminating information about pupils' performance, especially in relation to age differences, gender differences and differences within 20 percent attainment bands. The mathematics team has developed extensive coding of error responses to individual items, so the contrasts in performance relate to error and omission rates as well as facilities.

This information is included in short reports written specially for teachers. These have taken two forms: one is booklets on topics such as Decimals (Mason & Ruddock, 1986), Practical Mathematics (Foxman, 1987), and Attitudes and Gender Differences (Joffe & Foxman, 1988), and the Cockcroft Foundation List (Ruddock, 1988) and 4-page leaflets, mainly for primary teachers which highlight the main findings in particular areas. It is likely that much of the reporting of the 1987 results will have implications for teachers in mind.

Table 2
Comparing Decimals: Different Success Rates

Item 1 Which of the numbers below has the greatest value?	% of pupils selecting response	Item 2 Which of the numbers below has the smallest value?	% of pupils selecting response
A. 0.075	1	A. 0.625	34
B. 0.09	1	B. 0.25	3
C. 0.1	82	C. 0.375	2
D. 0.089	14	D. 0.125	37
		E. 0.5	22
Other	1	Other	1
Omit	1	Omit	1

Results

Initial analyses of most of the data from the survey of 11-year-olds in 1987 have been carried out, but those from the age 15 survey are still incomplete. The examples given below from 1987 therefore relate to the younger pupils only, in particular the discussion on calculator skills. Other examples relate to the APU's methods of presenting results, especially in documents for teachers. Since the overall pattern of results in 1987 is very similar to that of 1978 to 1982, some results from these surveys are included in the illustrations. In order to highlight factors which influence success, error and omission rates, there has been an emphasis on comparing the results of parallel items within and between assessment modes.

Concepts and Skills

Decimals. A feature of successive surveys has been the development of items to explore or extend particular findings obtained in previous surveys. For example, in the first age 15 survey in 1978 two written test items requiring pupils to compare decimals less than one obtained markedly different success rates. In response to Item 1 (Table 2) 82 percent correctly chose 0.1 as the largest decimal while only 37 percent were successful in selecting the smallest decimal in Item 2. Furthermore, while there were two popular responses to Item 1, there were three to Item 2. About 1000 pupils took each question illustrated in this section.

After experimenting with further items and interviewing pupils, these results were found to be due to two errors which we call "largest is smallest" or conversely, "smallest is largest" (LS error) and "decimal point ignored" (DPI error) respectively. For example, in Item 2 the LS error makers choose alternative A since this is the largest number after the decimal point; the DPI pupils select response E since this is the smallest number if the decimal points are ignored. The LS error was found to be largely unknown by teachers and other mathematics educators despite its high incidence at both age groups.

In Table 3, the results for Item 2 are contrasted with those for 2 new items, Item 3 and Item 4. Item 3 is the same as Item 2 except that an additional digit has been added to alternative C. Item 4 differs from Item 2 in that the largest instead of the smallest decimal is required. These changes are sufficient to make dramatic differences to the results. The correct and the two error responses are marked appropriately on the items in Table 3.

In the case of Item 4, those pupils responding correctly and those who make the DPI error select the same response. The reason for Item 1 obtaining such a high facility is now seen to be due to the correct response being also selected by those pupils making the LS error.

The surveys began in 1978 with totally separate Concepts and Skills item collections for the two age groups. When the results of the first surveys became known it was clear that there was a considerable overlap in performance between them and so in later surveys, an increasing number of items has been common to them both. In 1987 Item 2, above, was included in the age 11 survey together with a parallel version placing the numbers in a context. The out of context results are very similar to other items of this type in previous age 11 surveys.

These results show that the younger pupils' success rate is about 25 percent below that of the 15-year-olds, about average for the items common to both age groups. The effect of context on success rate is negligible, but

Table 3
Comparing decimals: The two main errors

Item 2			Item 3			Item 4		
Which of the numbers below has the smallest value?			Which of the numbers below has the smallest value?			Which of the numbers below has the largest value?		
			Percent of pupils			Selcting space		
A. 0.625	34%	(LS)	A. 0.625	4%		A. 0.625	60%	(Correct + DPI)
B. 0.25	3%		B. 0.25	2%		B. 0.25	0%	
C. 0.375	2%		C. 0.375	36%	(LS)	C. 0.375	0%	
D. 0.125	37%	(Correct)	D. 0.125	43%		D. 0.125	0%	
E. 0.5	22%	(DPI)	E. 0.5	13%	(DPI)	E. 0.5	33%	(LS)
Other	1%		Other	0%		Other	5%	
Omit	1%		Omit	2%		Omit	2%	

Table 4
1987 Age 11 Surveys: Comparing decimals in and out of context

Item 5			Item 6		
Which of the numbers below has the smallest value?			Which of the numbers below has the smallest value?		
A. 0.625	25%	(LS)	A. 0.625	12%	(LS)
B. 0.25	2%		B. 0.25	4%	
C. 0.375	1%		C. 0.375	2%	
D. 0.125	12%	Correct	D. 0.125	10%	Correct
E. 0.5	56%	(DPI)	E. 0.5	71%	(DPI)

there is a shift in the balance of the incidences of the two errors from LS to DPI. This was anticipated from the results of similar items for lower attaining 15-year-olds in another project carried out by the NFER for the DES (Foxman et al. 1988).

The results of the Concepts and Skills tests are also reported in five 20 percent attainment bands and they reveal another aspect of the decimal results.

For both items the LS error is made by more upper attainers than lower attainers, while the DPI error is a characteristic response of the lowest 40 percent. This was shown in previous surveys to be also the case for 15-year-olds. The LS error is, therefore, more "advanced" than the DPI error.

Calculator Skills in 1987

Tests of calculator skills have been used in all previous age 15 surveys in the 1-to-1 practical assessment. At age 11, the only calculator test was in the 1982 Primary Survey, so the picture of pupils' skills in this area has been considerably extended in 1987. Calculator use may be mandatory or optional. In the latter case calculators are made available for pupils to use, but it is up to them to decide if and when to use them.

The 1987 APU primary mathematics survey contained examples of both types. In written tests it is difficult to make calculators mandatory but it is possible in the 1-to-1 practical tests. There was one topic in the practical testing which was specifically concerned with calculator skills, and the assessor required the pupil to use the calculator to answer most of the questions put. The Calculator Skills topic included also the assessment of pupils' ability to approximate before calculating and to decide whether an answer was reasonable after doing so. In some other practical tests and in a number of the written Concepts and Skills, Theme and Problems and Patterns tests, calculators were available.

There are important differences between mandatory and optional uses of a calculator in what is being assessed. In a calculator available situation the ability to make an efficient choice between the calculator as the most effective way of reaching an answer and mental or pencil and paper methods is an integral part of the assessment. This means that pupils can avoid the use of a calculator in situations where they do not know how to use it for a particular calculation or do not know how to interpret the result in the calculator display. Such situations can be assessed in an interview where calculator use can be made mandatory.

Table 5
Attainment band analyses of decimals in and out of a context

	Ordering decimals: no context					Ordering decimals: in a context				
	Bottom 20%	Lower Middle 20%	Middle 20%	Upper Middle 20%	Top 20%	Bottom 20%	Lower Middle 20%	Middle 20%	Upper Middle 20%	Top 20%
Correct	1%	1%	4%	11%	49%	1%	1%	3%	5%	39%
LS Error	10%	12%	26%	46%	34%	2%	5%	13%	23%	21%
DPI Error	77%	81%	66%	37%	12%	84%	92%	80%	61%	33%

Two tasks, one from the practical tests and one from the written tests illustrate the difference in performance that was found to be associated with the calculator mandatory and calculator available modes.

Practical Topic

Calculator mandatory: "John spends £27.45 on shopping. His bus fare was 60p and a meal at a café cost £3.85. How much did he spend altogether?" (n = 310 pupils)

The correct answer, £31.90, was given by 28 percent of pupils, but the most common response (32 percent) consisted of variations on the digit string 91.3. Evidence on methods used by pupils have usually to be inferred from the answer given in a written test but a strength of practical assessment in the APU survey is the direct recording of method by the assessor present. For the question above, 20 percent of pupils worked in pence, and 37 percent entered 27.45 + 60 + 3.85 obtaining the digit string 91.3. In a written test when a calculator is available an item with smaller numbers produced a very different pattern of responses.

Calculator Skills Written Test

Calculator available: What is the cost of this shopping trip:

Bus fare 90p
Hamburger and coke £1.15
Shopping £4.25

(n = 392 pupils)

In this mode the correct answer £6.30 was given by 66 percent of pupils, with variations on 9.54 given by only 7 percent. The success rate was thus rather higher than those in the practical topics, and the proportion of pupils obtaining 95.4 on the calculator, by entering 90p as 90 rather than 0.9, and then using these digits in the answer given, 7% of pupils, is much lower. The difference between these results and those from the practical survey, where calculators are known to have been used, can be accounted for in several ways. Choosing not to use the calculator may be one factor, or using it and then rejecting the answer in favour of a later one obtained with or without the calculator is another.

These data support the view obtained from teacher ratings of frequency of calculator use that the British population of 11-year olds in 1987 was relatively naive in terms of calculator experience. When a calculator was mandatory for a calculation, as in the practical topics, mixed units were a large scale problem and interpreting the displayed answer (13.2) to £66 + 5 produced a range of responses such as 13.2 (39 percent of pupils), £13 and 2 pence (16 percent) and 132 (14 percent). Only 18 percent immediately responded with £13.20.

Dealing with mixed units and interpreting the display are essential calculator skills, but not widely mastered by 11 year olds in 1987.

Comparisons between response patterns to the same task in non-calculator and calculator available written tests were also carried out.

Assessment with and without a calculator for the same task: The apparently simple task of calculating 6.25–4 provides some interesting points for discussion.

Without a calculator the item tests both algorithmic competence and place value. When a calculator is used, the task should be a straightforward data entry exercise, but did not produce the success rate of over 90 percent expected from such tasks. Again, it seems likely that some pupils incorrectly judged that a non-calculator computation was the best method for them. The choice of when to opt for calculator-based computation rather than pencil and paper or mental working is a crucial one, and an aspect of calculator use which these data suggest needs further investigation. Items like the one above, which may appear deceptively easy, can be useful in this respect.

The same task given without a calculator being available produced a range of differences both in success rate and in response pattern. Apart from tasks which are straightforward data entry exercises with a

Table 6
Comparison of non-calculator and calculator available tasks

	Responses given					
6.25 – 4 =	2.25 (Correct)	225	6.21	621	Other	Omit
Non-calculator	31%	-	40%	3%	13%	13%
Calculator available	66%	4%	17%	2%	10%	1%

Non-calculator n=1000
Calculator available n= 392

calculator, but need awkward algorithms without one, two basic types were found:

- tasks where calculator use produces a higher success rate;
- tasks where calculator use produces similar or lower success rate.

Tasks where success rates were found not to be higher when a calculator was provided can be summarised as those where finding an appropriate method is the difficulty rather than the computational or algorithmic problems. For British 11-year olds such topics as rate and ratio and percentages showed this pattern.

Problem-Solving Strategies

Problem-solving strategies were tested in 1987 in the Problems and Patterns written tests, in some of the 1-to-1 practical topics, and in the small group problem solving. In the Problems and Patterns tests there are usually graded questions on five situations involving component problem-solving strategies, for example, continuing patterns, generalising them, and explaining how they work, working systematically, using trial and error methods, and so on.

In one example a subtraction is presented with two missing figures

$$\begin{array}{r} 5\square \\ -3\square \\ \hline 2\ 7 \end{array}$$

Pupils are first asked to supply one set of figures which will make the subtraction correct and then other possible answers. A similar problem follows which has six correct answers.

At age 11 in 1987, 22 percent of the pupils obtained all 6 correct answers, a slightly lower figure than that in the 1982 survey. In the 1982 survey of 15-year-olds 60 percent of them obtained all six correct answers. A higher proportion of the older pupils used systematic working to get their answers.

In general there is much more of a requirement for pupils to explain findings and to record their working than in the Concepts and Skills tests. In the above example pupils usually supplied sufficient evidence to judge whether their working was systematic, but in most situations it is very difficult to get them to record spontaneously. The advantage of the 1-to-1 practical tests is that pupils can be observed and can be asked about their methods of working. The 1-to-1 problem solving tests have included both mathematical situations and "everyday" problems such as arranging a Class Trip, organising a Birthday Party (11-year-olds), or Designing a Kitchen (15-year-olds).

With the advent of the small group assessment situation in 1987 the opportunity was taken to attempt some cross modal comparisons. One topic, Number Chains, was tried out in a written Problems and Patterns test and also in both the 1-to-1 and small group assessments. The situation involved applying a rule to whole numbers, "If its even, halve it; and if its odd, add 3". The effect of applying this rule to a number and then to the result of the transformation successively is to form a number chain e.g. 15 18 9 12 6 3 etc. All chains end in one of two loops, 6 3 or 4 2 1. In the 1-to-1 practical and the small group assessments the pupils were given plenty of opportunity to derive the rule themselves and to test out any conjectures they had made about what it could be before being presented with the substantive problem. This was to find out what sort of numbers end in 6 3 and what end in 4 2 1. The results show clearly the value of both pupil-teacher and pupil-pupil discussion in problems which are within the capacity of most 11-year olds to solve, as compared with a printed textbook presentation. Over a third obtained the correct answer in the small group assessment, with no help from an adult, compared with a quarter in the 1-to-1 who obtained it with a little or no help and only 1 percent in the written test version. In the small group an additional 13 percent got most of the way towards a correct solution without help and in the 1-to-1 practical a further 6 percent obtained the correct answer with a lot of help.

Attitudes and Gender Differences

In previous surveys pupils' attitudes have been studied by means of written questionnaires and by assessors' observations of pupils' responses to the 1-to-1 practical tests. In 1987 pupils' views were additionally sought on all written tests.

The attitude questionnaires investigate pupils' feelings towards mathematics generally and to individual mathematics topics on scales relating to enjoyment, utility, and difficulty. The results follow a similar pattern to that observed in other attitude surveys conducted both in the UK and in other countries: pupils find mathematics less enjoyable, more difficult, and less useful as they get older. More boys that girls perceived mathematics as being relevant to their futures, enjoyable and "one of their better subjects." Although boys and girls liked mathematics as a subject, the most frequently mentioned reason for disliking it was that it was too difficult; a reason provided by more girls than boys. More boys than girls thought that "...without maths our lives would be harder," that it is difficult to get on in life if "you haven't done much maths," and that it would help them to get a job one day.

In respect of performance, the surveys have consistently shown that gender differences across Concepts and Skills topic areas (computation-measurement top-

ics) are as great at age 11 as they are at age 15. They are also larger than the differences within topics which develop in favour of boys during secondary schooling. Most importantly of all, perhaps, is that nearly all the differences in performance between boys and girls are accounted for by the top 10 to 20 percent of attainers in most areas of mathematics at both ages. Thus, all the important gender differences are well established by the age of 11 in Britain.

Recently some more encouraging results for the girls have appeared. In the Problems and Patterns tests in 1982, girls were very slightly ahead at age 11 and even more so at age 15: a reversal of the trend obtained in Concepts and Skills.

Summary and Conclusions

The most important aspect of the project has been the development of new assessments and the breadth and richness of information provided by pupils' performance at the two age groups tested. A wide range of assessment modes has been employed including practical mathematics, problem solving in small groups, and mathematics on the micro. Calculator and mental skills have also been explored.

The 1987 age 11 survey results so far to hand show that the overall pattern of results is similar to that previously found in the phase of annual surveys from 1978 to 1982. In general, the Concepts and Skills and Problems and Patterns mean scores in 1987 are similar to 1982, but there are differences in detail: an upwards move in the mean scores of spatial sub-categories and one downwards in number sub-categories. There is evidence that these are probably linked to differences in emphasis in the mathematics curriculum, but that link has still to be established. An important feature of the work remaining will be to disseminate to teachers, by reports and other means, the results and their implications for teaching.

References

Foxman, D.D., Cresswell, M.J., Ward, M., Badger, M.E., Tuson, J.A. & Bloomfield B.A. (1980). *Mathematical Development. Primary Survey Report No. 1.* London:HMSO

Foxman, D.D., Badger, M.E., Martini, R.M. & Mitchell, P. (1981). *Mathematical Development. Secondary Survey Report* No. 2. London:HMSO

Foxman, D.D., Cresswell, M.J. & Badger. M.E. (1982). *Mathematical Development. Secondary Survey Report No. 3.* London:HMSO

Foxman, D.D., Martini, R.M. &Mitchell, P. (1982). *Mathematical Development. Secondary Survey Report No. 3.* London:HMSO

Foxman, D.D., Ruddock, G.J., Badger, M.E. & Martini, R.M. (1982). *Mathematical Development. Primary Survey Report No. 3* London:HMSO

Foxman, D.D., Ruddock, G.J., Joffe, L., Mason, K., Mitchell, P. & Sexton, B. (1985). *Mathematical Development. Review of the First Phase of Monitoring.* London:APU at DES.

Goldstein, H.G. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrica*, Vol. 73, pp.43-56.

Hutchison, D.A. & Schagen, I.P. (1987) Use of multilevel models in educational research. Paper presented at British Education Research Association Conference.

Mason, K. & Ruddock, G.J. (1986) *Decimals: Assessment at Age 11 and 15.* Windson:NFER-Nelson.

Ruddock, G.J. (1988) *Cockcroft Foundation List: The APU Results.* Windsor NFER - Nelson.

FINDINGS FROM THE FOURTH NATIONAL MATHEMATICS ASSESSMENT IN THE UNITED STATES

Jane O. Swafford · Edward A. Silver · Catherine A. Brown

Every four years, the National Assessment of Educational Progress (NAEP) gathers information about the mathematics performance of students in the United States at the elementary, middle, and high school levels. NAEP mathematics assessments were conducted during the school years ending in 1973, 1978, and 1982. The fourth and most recent assessment was conducted in 1986 with a nationally representative sample of approximately 35,000 9-year old, 13-year old, and 17-year old students.

The objectives that guided the development of the fourth mathematics assessment covered seven broad content areas:

1. Fundamental Methods of Mathematics
2. Discrete Mathematics
3. Data Organization and Interpretation
4. Measurement
5. Geometry
6. Relations, Functions, and Algebraic Expressions
7. Numbers and Operations

Each of these content areas was assessed at five process levels: understanding/comprehension, knowledge, skill, routine application, and problem solving.

From its inception, NAEP has developed assessments through a consensus process. The objectives that provided a framework for the fourth mathematics assessment were written and reviewed by a panel of mathematics educators, including classroom teachers. The objectives focused on content that should have been covered by a majority of students at a given grade level.

The first three assessments were conducted by the Education Commission of the States, whereas the fourth mathematics assessment was conducted by the Educational Testing Service. Using the framework of content categories and process levels outlined above, a group of mathematics educators worked with the staff of the Educational Testing Service to develop the items for the fourth mathematics assessment. The items were extensively reviewed by subject-matter and measurement specialists. A set of unreleased items from previous assessments was included in the fourth assessment to provide continuity and to establish a basis for measuring change in performance from previous assessments. The items were field- tested, revised, and administered to a stratified, multi-stage probability sample.

Some changes in methodology accompanied the change to the Educational Testing Service as the administrator of the assessment. In the fourth assessment, subjects were selected by grade level rather than by age. Matrix sampling procedures were used to identify a representative national sample of third-grade, seventh-grade, and eleventh-grade students.

There were also some changes in the actual administration of the assessment. Items from the previous assessment frequently were open-ended. The newly developed items were all multiple choice. In previous assessments, a paced audio recording was used to read each item to the students. Thus, the time allotted for each item was controlled. In this assessment, test items were divided into blocks of approximately 15 minutes each. Each student was administered a booklet containing three blocks of cognitive items and a six-minute background questionnaire. In order to provide broad coverage of topics, item-sampling procedures were used as in earlier assessments. Each student received approximately 10 to 15 percent of the items administered at each grade level. Approximately 2,000 students started each block of items, but because of the time limit some students did not complete all the items. Thus, performance on individual items that appear toward the end of a block is more difficult to interpret.

Previous mathematics assessments results have been reported on an item-by-item basis which has proven particularly useful to researchers, curriculum developers, and teachers. This practice was continued for the fourth assessment. The item-level results appear in companion articles in the *Arithmetic Teacher* (Kouba et al., 1988a, 1988b) and in the *Mathematics Teacher* (Brown et al., 1988a, 1988b) and in a monograph published by the National Council of Teachers of Mathematics (Lindquist et al. in press). Although item-level reporting of results has allowed for a reasonably clear and detailed description of the current level of performance of students in the United States, the level of detail makes it difficult to draw broad general conclusions about overall performance and how it has changed over time. In the past, NAEP has attempted to provide some summary of the data by aggregating them over

content areas or process levels at each grade level. This procedure has not proved to be entirely satisfactory because the aggregated scores have little meaning. For the fourth assessment, NAEP constructed scales in order to provide a profile of performance trends.

Performance Scales

In order to report performance trends across the assessments, a supplemental sample of subjects was selected by age rather than grade level and administered previously assessed items according to the procedures used in prior assessments. The item pool consisted of all items given in 1986 and in at least one of the previous two assessments. The total number of items included in the trends assessment was 68 items for 9-year olds, 98 for 13-year olds, and 94 for 17-year olds. The responses were scored, weighted in accordance with the population structure and adjusted for nonresponse. Item Response Theory (IRT) technology was used to estimate levels of mathematics achievement for the nation and for various subpopulations along a single scale.

With IRT it is possible to summarize the performance of a sample of students on a single scale, even if different students were administered different items. Using scaling techniques, NAEP was able to identify items that had similar statistical properties and use those items to define different levels of performance. These levels of performance were then used to generate, for different assessments and for different subpopulations, numerical scores that have criterion-referenced interpretations. That is, certain scores can be related to the attainment of certain skills on a hypothesized continuum of proficiency, and these scores can be used to describe the performance of different ages and subgroups based on a common standard. A more complete description of the scaling procedures can be found in *The Mathematics Report Card* (Dossey, Mullis, Lindquist, & Chambers, 1988).

IRT scales have a linear indeterminacy which may be resolved by an arbitrary choice of the origin and unit-size in each given subscale. The mathematics scale was linearly transformed so that the final scale would have

Table 1
Levels of Mathematical Proficiency

Level 150 - Simple Arithmetic Facts

Learners at this level know some basic addition and subtraction facts and can add two-digit numbers without regrouping. They recognize simple situations in which addition and subtraction apply. They also are developing rudimentary classification skills.

Level 200 - Beginning Skills and Understanding

Learners at this level have considerable understanding of two-digit numbers. They can add two-digit numbers, but are still developing an ability to regroup in subtraction. They know relations among coins, can read information from charts and graphs, and use simple measurement instruments. They are developing some reasoning skills.

Level 250 - Basic Operations and Problem Solving

Learners at this level have an initial understanding of the four basic operations. They are able to add and subtract whole numbers and apply these skills to one-step word problems and money situations. In multiplication, they can find the product of a two-digit and a one-digit number. They can also compare information from graphs and charts and are developing an ability to analyze simple logical relations.

Level 300 - Moderately Complex Procedures and Reasoning

Learners at this level are developing an understanding of number systems. They can compute with decimals, simple fractions, and commonly encountered percents. They can identify geometric figures, measure lengths and angles, and calculate areas of rectangles. These students are also able to interpret simple inequalities, evaluate formulas, and solve simple linear equations. They can find averages, make decisions on information drawn from graphs, and use logical reasoning to solve problems. They are developing the skills to operate with signed numbers, exponents, and square roots.

Level 350 - Multi-step Problem Solving and Algebra

Learners at this level can apply a range of reasoning skills to solve multi-step problems. They can solve routine problems involving fractions and percents, recognize properties of basic geometric figures, and work with exponents and square roots. They can solve a variety of two-step problems using variables, identify equivalent algebraic expressions, and solve linear equations and inequalities. They are developing an understanding of functions and coordinate systems.

Table 2
Percent of Students at or Above the Five Proficiency Levels

Proficiency Level	Age		
	9	13	17
150	98	100	100
200	74	99	99
250	21	73	96
300	1	16	51
350	0	0	6

Table 3
Average mathematics proficiency levels: 1973-1986

Proficiency Level Age	1973	1978	1982	1986
9	(219.1)	218.6(0.8)*	219.0(1.1)	221.7(1.0)
13	(266.0)	264.1(1.1)*	268.6(1.1)	269.0(1.2)
17	(304.4)	300.4(0.9)	298.5(0.9)*	302.0(0.9)

*Statistically significant difference from 1986 at the 0.05 level.
Jackknifed standard errors are presented in parentheses.

a weighted mean of 250.5 and a weighted standard deviation of 50 across all students in the three ages. An additional benefit of IRT methodology is that it provides for a criterion-referenced interpretation of levels on this continuum of proficiency. Although the proficiency scale ranges from 0 to 500, few items fell at the ends of the continuum. Thus, five levels of proficiency, ranging from 150 to 350, were chosen for describing the results. Each level is defined by describing the types of mathematics questions that most students attaining that proficiency level would be able to perform successfully. The levels are described in Table 1. The estimated proportion of each age level at or above each of the five proficiency levels is reported in Table 2.

It is important to note that since the items in the NAEP pool were not developed to conform to some hypothesized framework of levels of mathematical proficiency, and since the proficiency levels were derived in a *post hoc* analysis of performance, these levels do not represent an idealized picture of mathematical proficiency. Further, it is certainly possible to define hypothetical levels of mathematical proficiency beyond those identified here. The reported levels are merely those that emerged from the statistical scaling of the available pool of items. Caution is urged in interpreting the results based on these levels. Although statistically coherent, the available items do not necessarily fall into clearly defined clusters of related items.

National Trends

Average mathematics proficiency levels for each age group for the four mathematics assessments are given in Table 3. Significant gains have been observed for all age levels over time. (The proficiency levels reported for 1973 reflects a rough estimate of extrapolated results based on previously reported NAEP data.)

Performance of 9-year olds, which had shown little change from 1973 to 1982, improved significantly between 1982 and 1986. Performance of 13-year olds, which had increased in the late 1970's and early 1980's, leveled off between the last two assessments, registering virtually no change from 1982 and 1986. For 17-year olds, the downward trend that had been characteristic of performance in the 1970's was reversed. Seventeen-year olds made significant gains between 1982 and 1986. The gains for 17-year olds parallel the gains that were made by the same age cohort group between 1978 and 1982. Although the same cohort pattern is not reflected in the results for 9-year olds, the relationship at ages 13 and 17 suggest that the causes underlying the recent improvements at age 17 extend beyond recent reforms being made in high school graduation requirements. These performance trends as depicted by NAEP are pictured in Figure 1.

Trends among Minorities

Over the last decade, Black and Hispanic students have made significant gains in achievement in mathematics at all grade levels. Black students at all three ages have shown steady and significant gains across the past three assessments. Hispanic students at ages 9 and 17 have shown steady improvement over the past three assessments. At age 13, there was little change in performance between 1973 and 1978; however performance improved significantly from 1978 to 1986. In general, the gains of Black and Hispanic students have been greater and more consistent than the gains shown by White students. Nevertheless, although the gap between the performance of White students and the performance of Black and Hispanic students is narrowing, performance differences among these minority subpopulations remain significant at all three age levels. The gains suggest that programs implemented over the last ten to twenty years to improve the performance of minority students are having an effect, but even greater efforts are needed to provide real equity of educational opportunity for all American students.

Trends by Gender

Previous assessments have found few gender-related differences in mathematics achievement at ages 9 and 13, but at age 17, there have been small yet significant differences with males scoring higher than females. The same pattern occurred in 1986. Although

females. The same pattern occurred in 1986. Although there were no achievement differences at the youngest age level, more males than females obtained a proficiency level at or above Level 250 among 13- and 17-year olds. Differences were particularly evident among 13-year olds at Level 300, and among 17-year olds at Levels 300 and 350.

NAEP results also showed a significant advantage for males on geometry and measurement at Grades 3 and 11. Females tended to outperform males in the area of knowledge and skills while males showed a consistent advantage in the area of higher-level applications. There were no gender differences on the algebra subscale. As early as age 13, significantly more males than females responded that they were likely to enter a career that used mathematics, and more males than females responded that they were good at mathematics.

Performance Patterns

More critical than the changes in students' performance over time are the patterns of current achievement. Only 21 percent of the 9-year olds mastered basic mathematical operations and beginning problem solving skills (Level 250) that are usually taught in elementary school. One-fourth of them failed to demonstrate even beginning skills and understanding characterized by the next lower level of proficiency (Level 200). Among the 13-year olds, only 16 percent demonstrated a grasp of moderately complex mathematical procedures and reasoning (Level 300) generally embedded throughout the middle and junior high school curriculum in the United States. About one-half of the 17-year olds reached this level which can be characterized as being able to use moderately complex numerical procedures and to interpret simple inequalities, evaluate formulas, and solve simple linear equations. Less than 7 percent of the 17-year olds displayed abilities in multi-step problem solving and algebra. Closer examination of the results reveals that most of the progress that has occurred over the past eight years is in the domain of lower-order skills.

Overall, third-grade students performed well on selected whole number computation items but many appeared to lack mastery of place value and seemed to be learning mathematical skills at a rote manipulation level. About one-third of the seventh-grade students and one-fourth of the eleventh-grade students demonstrated extremely limited knowledge of some of the most basic mathematical concepts and skills. Although they could perform simple whole number calculations, they gave little evidence of knowledge of the most fundamental concepts of fractions, decimals, or percents. Similarly, they could identify simple geometric figures, make simple measurements, and read simple graphs, but they could not use basic properties of geometric figures, compute areas or volumes, or draw conclusions from graphs and tables. They lacked the ability to apply what they knew to a problem solving situation. At a time when mathematical skills are in high demand in the work place, few students in the last years of secondary school have mastered the fundamentals needed to perform more advanced mathematical operations.

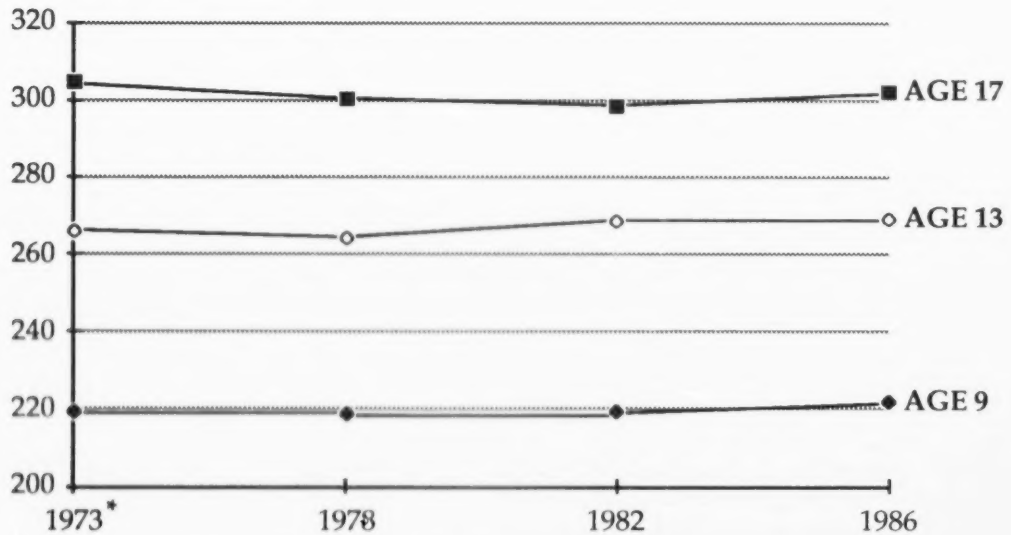


Figure 1. National trends in mathematical proficiency.

Learning Concepts and Skills

One of the central issues driving recent reforms in the U.S. mathematics curriculum has been the relative emphasis that should be placed on developing understanding of basic concepts and the teaching of mathematical skills. The modern mathematics movement in the United States in the 1960's emphasized understanding, whereas the "back-to-basics" movement which followed in the 1970's focused on teaching skills. It is not a question, however, of choosing between understanding and skills. There is mounting evidence that students cannot learn skills effectively in isolation but must understand the skills they are learning if they are going to retain them and be able to apply them in unfamiliar contexts.

The results of the fourth NAEP mathematics assessment suggest that many students are failing to develop an understanding of important concepts underlying the skills they are attempting to learn. The difficulty that third-grade students encounter adding three-digit whole numbers can be traced to their lack of understanding of place value concepts for three-digit numbers and older students' difficulties with fractions, decimals, and percents reflect serious gaps in their knowledge of basic fraction, decimal, and percent concepts.

Performance on the items in Table 4 illustrates how limited many students' knowledge is of the basic meanings of fractions and decimals. Many students who were successful at routine, frequently encountered calculations had difficulty when they were asked questions that did not involve standard calculations pre-

Table 4
Performance on Basic Number Concept Items

Item	Percent Correct Grade 3	Grade 7	Grade 11
A. What is 100 more than 498?	37	64	
B. 5 1/4 is the same as 5 + 1/4 5 – 1/4 5 x 1/4 5 ÷ 1/4		47	44
C. Write .037 as a fraction		48	58

Table 5
Generalizing the Formula for the Area of a Rectangle

Item	Percent Correct Grade 7	Grade 11
A. What is the area of this rectangle?	46	70
B. What is the area of this square?	13	45

Table 6
Performance on Algebra Items

Item	Percent Correct Algebra 1	Algebra 2
A. Solve: $6x + 5 = 4x + 7$	83	91
B. Simplify: $9(1 + 5x) + 13$	74	87
C. $x - y > x + y$ implies $y < 0$ $x > 0$ $x = 0$ $x = y$	38	50

sented in a familiar context, even when the questions involved basic number concepts.

The items in Table 5 illustrate the difficulty that students had generalizing the procedures that they had learned. Almost one-half of the seventh-grade students could calculate the area of a rectangle, but only 13 percent of them could apply this knowledge to find the area of a square, even though almost all students knew that the sides of a square are equal.

The items in Table 6 offer another example of the procedural orientation of many students. The majority of eleventh grade students who had completed one or two years of algebra could perform the symbolic manipulations involved in solving equations or simplifying expressions. Very few of them, however, could identify the relationships between variables that were implied by the equation in the third item.

Problem Solving

Although students at all three age levels could solve simple one-step word problems, they experienced difficulty with any nonroutine problem that could not be solved by the simple application of a familiar procedure. The results summarized in Table 7 illustrate the difficulty that students had with problems involving several steps. Most third grade students could solve the one-step problem in Item A. In fact, performance on this item was comparable to performance on similar computation items. There was a significant drop in performance on the two-step problem in Item B in spite of the fact that 85 percent of the third grade students could perform the additional computation required for this problem.

The pair of problems in Table 8 contrast eleventh-grade students' performance between a problem in which a standard procedure is sufficient and one in which understanding of the concept of average is needed.

For each of the previous NAEP mathematics assessments, performance on items assessing students' problem-solving abilities has provided the greatest cause for concern in the United States. This concern continues with the poor performance on multi-step problems and on nonroutine problems in the fourth assessment.

Instructional Indicators

In addition to administering items that measure student achievement in specific areas of the mathematics curriculum, NAEP gathers data that might be more generally considered indicators of instructional activity in school mathematics.

Table 7
Performance on Problem Solving Items

Item	Percent Correct Grade 3	Grade 7
A. Robert spends 94 cents. How much change should he get back from $1.00?	68	—
B. Chris buys a pencil for 35¢ and a soda for 59¢. How much change does she get back from $1.00?	29	77

Table 8
Applications of Skills

Item	Percent Correct Grade 11
A. Here are the ages of six children: 13, 10, 8, 5, 3, 3 What is the average age of these children?	72
B. Edith has an average (mean) score of 80 on five tests. What score does she need on the next test to raise her average to 81?	24

Table 9
Performance on the Same Items With and Without Calculators

Grade	Number of Items	Percentage of Items Correct With Calculator	Without Calculator
3	11	69	51
7	30	61	48
11	32	75	67

Mathematics Course Enrollment

At the time of the fourth NAEP mathematics assessment, over three-fourths of the 17-year old students in the sample reported that they were currently enrolled in a mathematics course. Moreover, the general trend appears to be toward taking more advanced mathematics courses. After previous declines, reported enrollments in Algebra II and more advanced mathematics courses (e.g., pre-calculus and calculus) increased between 1982 and 1986. Despite the increase, however, the data indicate that over 50 percent of 17-year olds had not enrolled in Algebra II and almost 40 percent of this age group reported not having taken any

mathematics course beyond Algebra I. In recent years, many states have increased the amount of mathematics required for graduation from secondary school. The proportion of students taking the more advanced mathematics courses, however, remains considerably less than 10 percent.

Classroom Instructional Activities

The fourth NAEP mathematics assessment included a variety of student background questions about the types of instruction in mathematics classes. For students at all three grade levels, typical mathematics instruction apparently consists of listening to teacher explanations, watching a teacher work problems at the board, using a mathematics textbook, and working problems presented on worksheets. About two-thirds of the seventh-grade students and over one-half of the third-grade and eleventh-grade students reported never working in small groups to solve mathematics problems. Although students reported a strong likelihood of working alone in mathematics class, approximately 80 percent of the students at all three grade levels reported that they never work on independent projects or laboratory activities in mathematics class. In general, these recent data suggest that little has changed in U.S. school mathematics instruction over the past decade.

Technology

The rapid growth in the general U.S. culture of available technological tools suggests that a parallel growth would have been seen in the Nation's schools. Some data from the fourth NAEP mathematics assessment provide glimpses of the extent to which technology has had an impact on mathematics instruction in the United States.

Almost all students at the three grade levels reported having access to a calculator at home. But relatively few reported having a calculator made available for use in school in mathematics class. In fact, about two-thirds of the seventh-grade students and approximately one-half of the third-grade and eleventh-grade students reported never having used a calculator (even their own) in mathematics class. When they were used in mathematics class, calculators were reportedly used most frequently to check answers. These data suggest that schools are lagging far behind the rest of American society in making available calculation tools and utilizing their potential for instruction and learning.

In addition to asking students about their use of the calculator for various tasks, a common set of problems was given to two equivalent samples of students, one sample using calculators and the other not. Students using calculators consistently performed better than students without calculators at all three grade levels. The difference, however, diminished with age. Their relative performances are given in Table 9.

Although students did better on the straightforward computation items given in the calculator assessment when calculators were available, overall performance on items for which calculators were available declined significantly for the two younger groups across the past three assessments.

The data on computer use and impact are somewhat more encouraging. Nearly one-half of the 13-year olds and more than one-half of the 17-year olds reported having access to computers to learn mathematics. This represents a major increase over previous assessments. It is not clear from the data how often students have access to computers or for what purpose. Nevertheless, it is clear that computers are increasingly available in U.S. schools for use in mathematics and that they are being used, at least sometimes, to enhance students' mathematical problem-solving activities.

Conclusion

Following the broad declines in student achievement that characterized the 1970's, it appears that there has been an upturn in achievement in mathematics in the United States in the 1980's. This trend provides little cause for complacency, however, as most of the progress occurred in the domain of lower-order skills. Student achievement at all age levels showed serious deficiencies. The discrepancy between students' desired and actual level of mathematics proficiency begins early on in schooling, and increases as they move into the upper grades. For minority students whose mathematics performance has tended to lie below national averages in NAEP assessments, the discrepancy between expected and actual performance for all age groups remains even larger than that for the nation as a whole, despite considerable gains since the last assessment.

The indications of a general increase in participation in advanced mathematics coursework is cause for hope for increased mathematical proficiency in the future. However, the emphasis on computational skills that generally characterizes school mathematics in the United States has left many students with serious gaps in their knowledge of basic underlying concepts. These deficiencies prevent students both from flexibly applying their knowledge and skills and from learning more advanced knowledge and procedures. Moreover, many of the skills that they have learned are in danger of becoming obsolete as technological advances alter the mathematics that adults need to function productively in society.

The curriculum reforms proposed by the National Council of Teachers of Mathematics (1987) in their recently released *Curriculum and Evaluation Standards*

for Schools Mathematics (draft) call for a reorientation of the school mathematics curriculum to place greater emphasis on helping students to become mathematical problem solvers and to communicate and reason mathematically. The results of the fourth NAEP mathematics assessment indicate that these are areas most critically in need of reform. Narrowing the gap between the current state of student achievement and classroom instruction and what should be constitutes a major challenge for American education.

REFERENCES

Brown, C.A., Carpenter, T.P., Kouba, V.A. Lindquist, M.M., Silver, E.A., & Swafford, J.O. (1988a). Secondary school results for the fourth NAEP mathematics assessment: Discrete mathematics, data organization and interpretation, measurement, number and operations. *Mathematics Teacher, 81* (4), 241-248.

Brown, C.A., Carpenter, T.P., Kouba, V.A., Lindquist, M.M., Silver, E.A., & Swafford, J.O. (1988b). Secondary school results for the fourth NAEP mathematics assessment: Algebra, geometry, mathematical methods, and attitudes. *Mathematics Teacher, 81* (5), 337-347,397.

Dossey, J.A., Mullis, I.V.S., Lindquist, M.M., & Chambers, D.L. (1988). *The mathematics report card: Are we measuring up?* Princeton, NJ: Educational Testing Service.

Kouba, V.L., Brown, C.A., Carpenter, T.P., Lindquist, M.M., Silver, E.A., & Swafford, J.O. (1988a). Results of the fourth NAEP assessment of mathematics: Number, operations, and word problems. *Arithmetic Teacher, 35* (8), 14-19.

Kouba, V.L., Brown, C.A., Carpenter, T.P., Lindquist, M.M., Silver, E.A., & Swafford, J.O. (1988b). Results of the fourth NAEP assessment of mathematics: Measurement, geometry, data interpretation, attitudes, and other topics. *Arithmetic Teacher, 35* (89), 10-16.

Lindquist, M.M., Brown, C.A., Carpenter, T.P., Kouba, V.A., Silver, E.A., & Swafford, J.O. (Eds.) (in press). *Results of the fourth mathematics assessment: National Assessment of Educational Progress*. Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics. (1988) *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Part 4

Selected Topics in Evaluation

THE GRADED ASSESSMENT IN MATHEMATICS PROJECT

Margaret Brown

The Graded Assessment in Mathematics project, which forms a classroom-based scheme of continuous and progressive assessment for students aged 11-16, is described in this paper. Successes and problems are briefly outlined, along with plans for future development.

Aims

The Graded Assessment in Mathematics (GAIM) project has as its aim the production of a continuous assessment scheme which will record the mathematical progress between the ages of 11 and 16 of students across the whole range of attainment. The intention is to take a broadly constructivist stance and thus focus on what the student understands and can use at a particular point in time rather than on what has been taught. The scheme has therefore been designed to be implemented alongside a variety of teaching schemes and teaching organisations, although an assumption has been made that the mathematics curriculum follows the broad principles laid down in the Cockcroft Report (Department of Educational Science (DES),1982).

At a more detailed level, the aims of the Graded Assessment in Mathematics project include:

1. To provide an explicit, continuously updated record of the mathematics that students know, understand, and can apply, in order to:

- help teachers better match the curriculum to the student;
- help students become aware of their progress, and more actively involved in their own learning;
- provide information for parents, heads, colleges, employers, in whatever degree of detail is desired.

2. To encourage a curriculum which conforms to the recommendations of the Cockcroft Report and in particular includes:

- investigations and practical problem-solving;
- discussion, group work, and extended work;
- a focus on process as well as content;
- emphasis on understanding and applying concepts, rather than on knowledge of specific techniques;
- a broad range of mathematical ideas.

In addition to these educational aims, in order to be attractive to teachers it was necessary to have a third, more pragmatic aim:

3. To link into other assessment and curriculum schemes by the provision of:

- the facility to convert the continuous record into a summative grade which is accepted as valid without any supplementary final examination in the General Certificate of Secondary Education (GCSE), the new national examination at age 16+;
- graded assessment profile certificates from one of the national examination groups which will feed in to the Record of Achievement schemes at present being trialled as a feature of government policy (DES 1984,1987a);
- specific guidance for teachers who wish to integrate the GAIM assessment scheme with the most popular published curriculum schemes;
- a close match with the structure of the proposed national curriculum (DES 1987, 1988a,b) which will enable GAIM to be used as the teacher assessed element of the national asessment at 7, 11, 14, and 16.

Background

The development of "graded tests" which more recently have evolved into "graded assessments" constitutes a significant innovation in classroom assessment. It has taken place in England during the last decade, starting in the mid-nineteen-seventies in the area of modern languages (Harrison, 1982; Pennycuik & Murphy,1988).

Teachers of modern languages in England had experienced a problem of student drop-out which became particularly severe after the change to all-ability comprehensive schools. Feeling a need both to increase motivation and to provide evidence of achievement for students who did not continue long enough to sit the public examinations at 16+, a number of local teacher groups developed systems of graded tests, following the model of examinations for professional interpreters. The characteristics of graded tests were defined as:

- progressive, with short-term objectives leading on from one to the next;

- task-oriented, relating to the use of language for practical purposes;

- closely linked into the learning process, with pupils or students taking the tests when they are ready to pass.(Harrison, 1982)

In fact, most of the graded test systems shared two other common features. First, they were generally organised into a set of successive levels, designed in most cases so that the median student might expect to pass level 1 at the end of year 1 in the secondary school, level 2 at the end of the second year, and so on. Second, they were not only task-oriented, but also incorporated at each level a set of objectives (grade criteria). Thus most graded test systems were criterion-referenced, with criteria which tended to the active and the practical.

The first evaluation of a graded test system was extremely positive in terms of attitudes of students, teachers and parents (Buckby et al., 1981). Rises of the order of 20 percent in the numbers of pupils continuing with the study of modern languages were reported fairly consistently across different schools.

Not surprisingly, evidence of improvement in student motivation on this scale attracted attention among officers in local education authorities, two of which were to form consortia with examination boards in order to develop graded test systems in the major subjects.

The GAIM project is the mathematics scheme which forms part of the London consortium, in which the partners are the Inner London Education Authority, King's College London, and the London East Anglia Group for GCSE (which includes the University of London School Examinations Board). Within the consortium, parallel graded assessment schemes are under development in science, modern and community languages, English, and craft,design and technology. The GAIM project also receives generous funding from the Nuffield Foundation.

By 1983, approaches to assessment had become much broader, incorporating, for instance, the possibility of observation by teachers during classwork and assessment negotiated between pupil and teacher. The term "graded tests" was therefore superseded nationally by "graded assessment" so as to allow such non-test methods where appropriate.

Research Basis

The notion of levels of attainment in mathematics, which provides the basis for a graded assessment scheme, was the subject of a large-scale investigation of secondary students' understanding of mathematics, undertaken at King's College London (previously Chelsea College) in the nineteen-seventies (Hart, 1981; Hart, Brown & Kuchemann, 1985). Principal findings of this study, "Concepts in Secondary Mathematics and Science"(CSMS), were:

- in spite of the fact that students had often followed a similar curriculum, the range of attainment across any single age group was very large;

- progress from one year to another was relatively slow, particularly so in relation to the attainment range in any single age-group;

- many students were using only rather primitive mathematics, much of which had been taught in the infant school (ages 5-7);

- students rarely made use of methods taught at school, but preferred their own idiosyncratic methods, which were often specific to a particular problem and not generally applicable;

- within each topic (such as ratio, graphs, etc.) a series of from 4 to 7 levels of attainment could be differentiated, and students appeared to progress through these levels in a consistent way (i.e. even with 7 levels, not more than 7 percent of students appeared to have achieved one level without achieving all the levels below it).

The work of the CSMS study thus identified the low correlation between what the Second International Mathematics Study terms "the implemented curriculum" and "the attained curriculum". This highlights the need for an accurate record of what each individual student knows, understands, and can apply, in order to assist teachers in reducing the degree of curriculum mismatch.

The CSMS study also provides considerable data as to which mathematical concepts and skills might be included at each level of a graded assessment scheme, once the definition of the levels is agreed. Other information used to assist in this task derived from the results of other research projects at King's College (Dickson, Brown and Gibson, 1984; Booth, 1984; Hart, 1984; Kerslake, 1986; Denvir and Brown, 1986; Hart et al, in press). Further survey data was available from the Assessment of Performance Unit (DES, 1986) and from the examination boards.

GAIM Structure: Levels

It was necessary near the start of the GAIM project to determine the number of levels to be included in the scheme. The precedent of about one

level per year which had been set by the modern languages graded test schemes seemed to have proved satisfactory; it was felt that the result of any fewer levels than this would be that students would become discouraged. The results of a government-funded evaluation study of a graded test scheme for low attainers in mathematics later confirmed this (Close & Brown, 1988).

However, in contrast to the study of modern languages, which normally begins in England at age 11, there was considerable evidence that students were already at many different levels in mathematics at the start of secondary school. The Cockcroft Report (DES, 1982), drawing partly on CSMS data, suggests a range of seven years of achievement for students at 11. In fact closer scrutiny of the CSMS data suggests that the range is at least 10 years, since while low attaining 11 year olds behave mathematically like 7 year olds, advanced 11 year olds are considerably ahead of average 14 year olds.

Hence it was decided that GAIM should have 15 levels, allowing for the brightest students at 11 to make a further 5 years' progress. Having studied data of examination grades in the previous national examinations at 16+, it seemed reasonable to identify the last seven GAIM levels with the seven grades of the GCSE since the median 16 year old previously obtained the grade which would be equivalent to level 10. The positioning of the boundaries for the earlier GAIM levels was done by reference to CSMS data, gathered on each year group from 11-15, so as to maximise the chances of students progressing at the rate of one level per year.

The result of fixing the levels in this way is that the earlier levels are "closer together" in mathematical terms than the later levels, since the students on later levels are expected to make greater progress in each year.

Although the earlier work at King's College (Hart, 1981; Denvir and Brown, 1986) does support the idea of hierarchies of learning, with the order in which conceptually based skills are developed relatively invariant within specific local branches, there is evidence that not all children progress uniformly across all mathematical areas. Also, the hierarchies may only hold for a particular education system at a particular time.

For these reasons, GAIM discourages teachers from teaching and assessing one level at a time, as was assumed in the earlier graded test model. Instead it is suggested that students' current records contain details of several adjacent levels, so that teachers may, on a particular topic, assess how high students can go in their attainments, even if this is at higher levels than those at which students are generally working.

The learning theory assumed is thus a constructivist one, in the simple if not in the radical sense (Kilpatrick, 1987) in which it is expected that children gradually construct their own mathematical knowledge in relation to the experiences they have had. They should not therefore be assessed only on the parts of the curriculum they have recently been presented with, as is the current custom. To give weight to this recognition of diversity in learning patterns, the students will receive customised profile certificates recording the highest level reached in each separate topic.

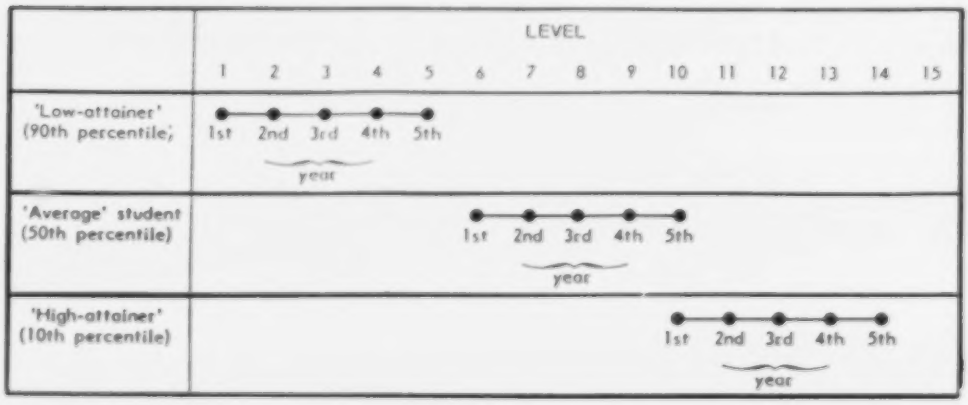

LEVEL
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
'Low-attainer' (90th percentile)
1st 2nd 3rd 4th 5th
year
'Average' student (50th percentile)
1st 2nd 3rd 4th 5th
year
'High-attainer' (10th percentile)
1st 2nd 3rd 4th 5th
year

Figure 1. The expected progress of students at the 10th, 50th and 90th percentiles respectively.

GAIM Structure: Content and Process

The GAIM assessment scheme is based on two components: topic criteria and coursework activities.

Topic criteria are a bank of profile statements, or objectives, which aim to describe the mathematics that students know, understand and can apply. They are organised into the 15 levels of difficulty described above, with about 20-35 criteria at each level, and into 6 topic areas: logic, measurement, number, space, statistics, algebra and functions.

Within each topic area the criteria are arranged within topic strands developing through the levels, enabling the structure of the scheme to be readily appreciated by teachers and students. Insofar as is possible, topic criteria relate to conceptually based skills and processes and not to rehearsed techniques.

Examples of topic criteria:

> Can give an answer to a general question, or decide when a conjecture is true, by producing and/or testing some specific examples. (Logic, level 3)

> Appreciates that multiplication by a number less than 1 reduces. (Number, level 11)

Coursework activities are open activities which are designed to encourage student decision-making. It was deemed necessary to incorporate such activities to stimulate the synthesis of skills, as without them the topic criteria might encourage fragmented teaching. Activities are suitable for a wide attainment range, allowing students to tackle the problem at their own level. They can be used with whole classes, with groups, or with individuals. There are two types of activity, investigations and practical problems, reflecting the pure and applied aspects of mathematics respectively.

Examples of investigations and practical problems:

> Investigating the different symmetry patterns obtained by shading squares on a grid.

> Investigating the number of ways of giving change for different sums of money.

> Planning the layout of newspaper advertisements.

> Scheduling the manufacture of garments in a small co-operative.

Assessment and Recording

At one level, the GAIM materials (GAIM team, 1988) can be used in any way a school wishes. For example, teachers may use the topic criteria to help write their own school profiling scheme, or the coursework activities alone to help assess the attainment level of the students. However, if schools want their students to receive certificates recording their performance from the examination group, and to use the scheme as an alternative way of gaining grades in the national GCSE examination at 16+, the school must be accredited by the examination group, and visited regularly by an external assessor to check that the school is carrying out assessment procedures properly, and is using equivalent standards to other schools.

Coursework activities are assessed according to the overall level of the work. Because of the degree of openness of the activity it is not possible to give precise instructions for marking; nevertheless, guidelines are provided for each task as to the sort of performance expected for each of the levels, together with diverse examples of students' work. In practice, after a short training session, teachers find these easy to use and a satisfactory degree of agreement is achieved. The teachers' notes receive extensive trialling before publication. It is intended that teachers experienced in using the scheme should be encouraged to use activities from other sources, and work is progressing on a more general set of guidelines for this.

Topic criteria should ideally be assessed as a result of the student's performance in open activities, since the fact that students can apply their knowledge in such situations provides reliable evidence for its acquisition. As part of the teacher's notes for each activity are listed those topic criteria that trialling has shown are most likely to be demonstrated.

In practice only a minority of criteria are assessed in this way, and the remainder have to be assessed during normal classwork. Because there is a wide range across different schools in teaching organisation and teaching material, it is not realistic to prescribe exactly how assessment of each criterion should be carried out. Hence in the end it must be left to the external assessor to check that procedures are appropriate.

To assist assessors, certain safeguards are built in. For example, written evidence must exist for at least 50 percent of the criteria for any student. A single written item alone however is not considered sufficient, since students are expected to be able to apply knowledge in different contexts. For the same reason, teachers are asked to ensure a delay of at least two weeks, and preferably more, between any direct teaching related to a criterion and student assessment.

This still gives teachers the opportunity to assess

some items orally and/or practically, without the requirement of written work by the student. Sets of criteria written so as to be comprehensible to students are provided to encourage students to record their own achievements and to volunteer to the teacher when they feel they have attained a criterion. The teacher however still needs to demand and judge the supporting evidence.

Although no formal evaluation of GAIM has yet been carried out, many teachers have reported increases in student motivation as a result of students more actively participating in their own assessment and being aware of short term goals which are perceived to be realisable. Teachers have also reported that they have gained much more insight into what students know and what they find difficult as a result of focussing on individual achievement in a detailed way, whereas previously their assessment related closely to their own teaching. Against this, teachers have found it to be a radical step which is quite onerous, particularly to begin with, in terms of workload.

The Development Programme

The GAIM project has been evolving over five years and has one more year of the development phase to go before the running of it is taken over entirely by the London East Anglia Group of examination boards. During that time hundreds of teachers have contributed to its development, some generously seconded full-time for one or two years by the Inner London Education Authority or other authorities, many with regular half day releases to attend feedback and development meetings, and others corresponding with us or attending occasional weekend or one-day conferences. Over seventy pilot schools, organised in clusters in twelve local authorities, will be working with the project from September 1988. This will be about 20 more schools than are currently involved.

A development package is now published with material for the first eight levels (GAIM team, 1988); the complete publication is planned for 1990. We are still working on new support materials to assist teachers in running the scheme, and will be closely monitoring the first awarding of GCSE certificates in 1989.

A major need will be to modify the scheme so that it will provide one form of the teacher-assessed component of the national assessment, due to begin operation for 14 year olds in 1991. This should be a reasonably simple task, since GAIM, along with the London-based graded assessment schemes in other subjects, is acknowledged to have been chosen as the model adopted, with some regrettable modifications to fit with government policy, for the national assessment scheme.

References

Booth,L. (1984). *Algebra: Children's Strategies and Errors*. London : NFER-Nelson.

Buckby,M., et al. (1981). *Graded Objectives and Tests for Modern Languages: An Evaluation*. London: Schools Council Publications.

Close,G. & Brown,M. (1988). *Graduated Assessment in Mathematics: Report of the SSCC Study*. London: Department of Education and Science.

Denvir,B. & Brown,M. (1986). Understanding of number concepts in low attaining 7-9 year olds: Part 1. Development of descriptive framework and diagnostic instrument. *Educational Studies in Mathematics* 17, 15-36.

Department of Education and Science, Assessment of Performance Unit (1986). *Mathematical Development: A Review of Monitoring in Mathematics, 1978 to 1982*. London: Department of Education and Science.

Department of Education and Science (1984). *Records of Achievement: A Statement of Policy*. London: Her Majesty's Stationery Office.

Department of Education and Science, Committee of Inquiry into the Teaching of Mathematics in Schools (1982). *Mathematics Counts ('The Cockcroft Report')*.London: Her Majesty's Stationery Office.

Department of Education and Science, National Curriculum Mathematics Working Group (1987b). *Interim Report*. London: Department of Education and Science.

Department of Education and Science, National Curriculum Mathematics Working Group (1988b). *Final Report*. London: Department of Education and Science.

Department of Education and Science, Records of Achievement National Steering Committee (1987a).*Records of Achievement: An Interim Report*. London: Her Majesty's Stationery Office.

Department of Education and Science, Task Group on Assessment and Testing (1988a). *A Report*. London: Department of Education and Science.

Dickson,L., Brown,M. & Gibson,O. (1984). *Children Learning Mathematics: A Teacher's Guide to Recent Research*. London: Cassell, for the Schools Council.

GAIM (Graded Assessment in Mathematics) Team (1988). *GAIM Development Pack*. London: Macmillan.

Harrison, A. (1982). *Review of Graded Tests: Schools Council Examination Bulletin 41*. London: Methuen Educational.

Hart, K. (Ed). (1981). *Children's Understanding of Mathematics: 11-16*. London: John Murray.

Hart,K. (1984). *Ratio: Children's Strategies and Errors*. London: NFER-Nelson.

Hart,K., Brown,M. & Kuchemann,D.E. (1985). *Chelsea Diagnostic Tests*. London: NFER:Nelson.

Hart,K., Johnson,D.C., Brown,M., Dickson,L., & Clarkson,R. (in press). *Children's Mathematical Frameworks 8-13: A Classroom Study*. London: NFER-Nelson.

Kerslake,D.(1986). *Fractions: Children's Strategies and Errors*. London: NFER Nelson.

Kilpatrick,J. (1987) What constructivism might be in mathematics education. In J.C.Bergeron et al. (Eds.) *Proceedings of the Eleventh International Conference on the Psychology of Mathematics Education* (Vol.1, pp 3-27). Montreal: Universite de Montreal.

Pennycuik,D. & Murphy,R. (1988) *The Impact of Graded Tests*. Basingstoke,U.K.: Falmer Press.

AN INFORMAL DIAGNOSTIC INSTRUMENT FOR ALGEBRA: RATIO AND PROPORTION

Douglas Edge

Collection of information-rich data is extremely important to diagnosticians and researchers. In mathematics education this is typically accomplished by asking students to complete either a pencil-and-paper test or by speaking with students using think-aloud or structured interview techniques. Each approach has its strengths and weaknesses. Pencil-and-paper tests permit large scale data collection yet cannot provide much more than superficial error-pattern analysis. Interview techniques facilitate collection of more detailed information, often focussing on conceptual development, but the time consuming nature of interviewing means that the number of students involved in such studies must remain relatively small. The major purpose of this study was to investigate whether or not a diagnostically oriented test could be developed which would incorporate the advantages of both approaches to data collection.

The topic chosen for this study was "concepts in early algebra" and focussed on students in Grades 7 and 8; that is, students in the two years immediately preceding entry into secondary school in Ontario. The test developed consisted of twelve questions: four to focus on students' notions of equation, four on variable, and four on ratio and proportion. This paper reports on the questions relating to ratio and proportion.

Background

Writing in mathematics classes is not new. Burton (1985) outlined different forms that such writing could take to promote the development of intellectual skills essential to the understanding of the discipline. Stempien and Borasi (1985), like Burton, focussed on writing as a learning tool by asking students to write mathematics-related stories, essays, and diaries. They concluded that writing provided opportunities for students to clarify their understanding of concepts and helped students organize their ideas.

Gordon (1988) investigated the use of writing strategies with students enrolled in developmental-studies algebra classes. He compared three classes where students had to write about their mathematics assignment with three other classes, which were either given extra exercises and or simply discussed previously assigned work. Although Gordon was cautious about attributing his findings to the differences in treatments, he acknowledged the value of the writing strategy. Bright (1988), also working with college level students, studied story editing as a methodology for identifying conceptual understanding in geometry. He found the story editing helpful in that it revealed otherwise undetected misconceptions.

Specific to testing, Ashlock (1987) pointed out that pencil-and-paper tests can be used to examine both skills and concepts, but that it is much more difficult to design items for tests that permit us to infer students' understanding of a concept. He provided examples of three different types of pencil-and-paper items that would be suitable for diagnosing conceptual understanding. The items involved sentence completion for ideas or rules, symbolization for statements made up of numerals and signs, and portrayals for drawings that model the concept in some way. Olson (1987) designed an algebra readiness assessment device that, like Ashlock's, used diagrams, comparisons, and so on, but also asked students to explain their choices. Olson's test included items related to class inclusion, transitivity, concept of equation, and proportional reasoning.

The major focus of this study was students' ability to explain their answers in written form. The specific insights into children's understanding of proportional reasoning were secondary to that focus.

Method

Ninety-seven students ranging in age from 12 years 4 months to 14 years 7 months in Grade 7 and 13 years 5 months to 16 years 7 months in Grade 8 participated in the study (see Table 1). This group represented all the students enrolled in the Grade 7 and 8 classes in an elementary school located in a town in rural southwestern Ontario. The testing occurred during June, the last month of the school year.

Table 1
Number, mean age in months (with standard deviation) and range of ages in months

		Grade 7	Grade 8
Male	Number	23	24
	Mean age	157.1 (6.3)	170(7.7)
	Range	148-175	161-199
Female	Number	30	20
	Mean age	155.8 (4.2)	166.9 (3.0)
	Range	150-167	161-173
Total	Number	53	44
	Mean age	156.4 (5.7)	168.6 (6.2)
	Range	148-175	161-199

Prior to administering the test, it was explained to the students that we were trying to find out why some students find mathematics easier to learn than others. "Unfortunately people don't come with zippers in the heads (an audible "yuck" was heard). So, we have designed a special test, different from those you usually see. Let's try a sample question together." Students were then asked to discuss which fraction was larger and why, 1/5 or 1/6 , and practice writing down their explanation. A few sample explanations were discussed. Finally students were asked to complete the test. "Try your best to help us. There is no time limit." No student took longer than 45 minutes for the twelve- item test.

Questionnaire

The last four items on the test were the ones related to proportional reasoning. Reference to these items will be done using numbers nine through twelve, the item numbers on the original test. See Figure 1 for the four items. Note that on the actual script the items were placed one question per half page. After the question was presented, a four-centimetre work space was given followed by the statement, "Please explain how you got your answer."

Item 9: If 8 tickets cost $12.00, how much would we have to pay for 6 tickets?
Please explain how you got your answer.

Item 10: Chris was asked to trim the trees in the gardens of three families. The Adams' garden had 10 trees; the Brown's had 15 trees; and the Campbell's had 25 trees. It took Chris 2 hours to trim the trees in the Adams' garden. How long would it take to trim those in the Brown's garden? How long would it take to do those in the Campbell's garden?
Please explain how you got your answers.

Item 11: We measured the heights of two rectangles with sticks. The height of the short one was 4 sticks. The tall one was 6 sticks. We measured the height of the short rectangle again, this time with loops. The short one was 6 loops. How many loops would we need for the height of the tall one?
Please explain how you got your answers.

Item 12: One flagpole, which is 8 metres high, casts a shadow 3 metres long. Another flagpole casts a shadow 5 metres long. How long is the second flagpole.
Please explain how you got your answer.

Figure 1. Proportional-reasoning items on the early-algebra assessment test.

For Item 11, a pair of rectangles, proportionally drawn, approximately two and three centimetres in height, were sketched to the left of the question. For Item 12, a pair of flag poles, again proportionally drawn and about two and three centimetres in height, were displayed to the right of the question.

The four proportional reasoning questions were not intended to represent a hierarchy of skills. Each question was selected because it either addressed a specific content objective or provided for some form of pedagogical insight.

Item 9 was a modification of Hart's (1984) recipe question which gave ingredients suitable for 8 persons then asked what would be needed for 4 and for 6. For Item 9 it was decided to provide information about 8 tickets but then directly ask for the information about 6 tickets. Would children opt for doubling and halving strategies? Item 10 was prepared to parallel Hart's "eels being fed sprats" problem. This question was selected because the ratio of the numbers was 2:3:5 and possibly would provide more variation in answers and explanations than had the ratio been easier, say 1:2:3. Item 11 was an adaptation of Karplus' "Mr. Tall, Mr. Short" question that Hart used in her Concepts in Secondary Mathematics and Science (CSMS) (Hart, 1981) and Strategies and Errors in Secondary Mathematics (SESM) (Hart, 1984) studies. Rectangles were used for this item because they could be drawn with the same width. Unlike the Mr. Tall/Mr. Short figures, they varied proportionally only in height. The numbers in the proportion 4:6 = 6:n were considered particularly good in that they could facilitate either additive or multiplicative solution strategies. Item 12 was written to incorporate a ratio where one term was not a multiple of the other. Students could not simply use doubling and halving, or factor techniques to find the height of the second flagpole. Would "difficult numbers" result in different strategies? Would students be able to explain the fractional aspects of the problem?

Analysis

Each item was scored using a two-part code. Answers were coded as 0 (incorrect), 1 (correct), or 2 (omitted). Explanations for each answer were also coded. These latter codes were developed as needed. For example, if on the first script scored, the student provided an explanation such as "I multiplied", that phrase would be assigned a 01 code. If on the second script, the explanation was "first I divided, then I multiplied", that would be assigned a 02. If the explanation on the third was "I multiplied", no new code would be needed. Examples of codes for Item 9 are: 1-01, correct answer followed by "divided 12 by 8 and got the $1.50, then I multiplied by 6 to get the answer"; and 0-06, incorrect answer followed by "I multiplied 6 by 12." For Item 9 a total of 18 explanation codes were needed; for Items 10, 11 and 12, the number of codes needed were 19, 10, and 22, respectively. These codes were then categorized and summarized. Coding, categorizing, and summarizing data in this manner facilitated further analysis in two ways: on a question-by-question basis across all students, and on a student-by-student basis across each set of items on any one script.

Results

Item 9: The analysis of data with the code summary for this item is presented in Table 2. Of those who answered this item correctly, 22 of 29 Grade 7s and 23 of 26 Grade 8s chose some form of unitary analysis. No student used the traditional unitary analysis form of writing equivalent sentences one beneath the other. Most showed 12÷8 = $1.50 and $1.50 x 6 = $9.00. For explanations, students typically stated that they had divided the 12 by 8, then multiplied the answer by 6. A variation of this procedure was done by two students who divided the 12 by 8 to obtain the $1.50 but then multiplied the $1.50 by 2, to equal $3, and then subtracted the $3 from the $12. A third student provided as explanation a kind of mathematical two- step: "I just subtracted and did a little division." In another variation of unitary analysis, three students used guess-and-check strategies indicating that they guessed at $1.00 for one ticket "which would be $8.00 total", so they "needed" an additional $0.50 per ticket. The other two students in this category gave variations of the same answer: "I guessed $2 for each one which was $16 so the answer had to be $1.50"; and "I just guessed $1.50 for each ticket and it happened to be right." On the incorrectly answered questions, the majority of students used the numbers in the problem but to perform inappropriate algorithms. Typical of these responses are $10 (I subtracted $2 from $12), $72 (I multiplied 6x12), and $2 (12÷6=2 so the cost is $2). The students' arithmetic explanations were clear. What is not clear is why they chose particular operations. The other major category of incorrect responses is "correct reasoning wrong answer." In all six cases the error occurred when dividing the $12 by 8 (for example, 12÷8 = $1.45, and 6 x $1.45 = $9.70).

Table 2
Analysis of data with code summary for Item 9

Code	Status/Explanation	Grade 7 (n=53)	Grade 8 (n=44)
2 Omitted		6 (11%)	9 (20%)
1 Correct		29 (55%)	26 (59%)
	unitary analysis stated	23	22
	answer a guess	4	1
	no explanation	2	2
	correct answer wrong reasoning	0	1
0 Incorrect		18 (34%)	9 (20%)
	Correct reasoning wrong answer	3	3
	wrong operation	13	3
	partially correct	2	1
	incomplete	0	2

Item 10: Analysis of data for this item is given in Table 3. Of the correct responses, 20 of 28 students in Grade 7 and 16 of 24 in Grade 8 reasoned that if it took 2 hours to trim 10 trees then it would take 1 hour for 5, hence 3 and 5 hours, respectively, for the Brown's and the Campbell's. Many in this group did not show any calculations. They wrote their answers and provided an appropriate explanation. Five students changed the hours to minutes, then wrote that it took 120 minutes for 10 trees so one tree would take 12 minutes, and so on. One student indicated that, "if Chris cuts the trees at the same rate, then he will do one and a half times the two hours for the Brown's trees." Two students focussed on multiples of five: "the numbers were multiples of five that could be reduced to lowest terms" and "10 is 2, 15 is 3, 20 is 4 and 25 is 5." For the incorrectly answered items, the most prevalent response is 2 1/2 hours for the Brown's, 41/2 hours for the Campbell's. Student #26, for example, wrote "it's 2 hours for 10 trees so 30 minutes is for 5, and that's why 15 is 21/2." In the incorrect explanation category, "operation/units confusion", one student wrote "You subtract 2 hours from 10 trees giving 8; so now you subtract 2 from 15 and 2 from 25 to get the answers of 13 and 23." Another student indicated that it would take "2 hours + 5 extra trees for 7 hours for the Brown's and 2 + 5 +10 = 17 for the Campbell's." A sample response in the "explanation unclear" category is "I know what I'm doing, but I just screwed up. Divide by 3, use percent and divide how many trees put into two hours and so on forget it just use division or percent."

Table 3
Analysis of data with code summary for item 10

Code	Explanation Category	Grade 7 (N=53)	Grade 8 (N=44)
2 Omitted		7 (13%)	11 (25%)
1 Correct		28 (53%)	24 (55%)
	found 5 trees per hour	20	16
	changed minutes to hours	1	4
	B=1 1/2 A, C=2 1/2 A	1	0
	multiples of 5's	1	1
	correct answer no explanation	2	2
	explanation unclear	1	1
0 Incorrect		18 (34%)	9 (20%)
	used 1/2 additively	10	3
	correct reasoning/ incorrect-answer	1	3
	operation/units confusion	2	2
	explanation unclear	3	1
	wrong answer/no explanation	2	0

Item 11: Refer to Table 4 for the data summary. The overwhelming response for this item was the incorrect category of "add 2". Students generally explained that the difference in sticks between the heights of the rectangles was two, so the difference in loops was also two. Hart (1981) labelled these students "adders" (p. 94). They tended to see an additive relationship in proportional thinking rather than the multiplicative one. In Hart's study about 32 percent of subjects were "adders". In this study the percent of adders was almost double this amount. Of the students who answered this item correctly, their explanations were clear and accurate: "Six loops and 4 sticks to the short rectangle would mean 11/2 loops to every stick which means 6 loops + 6 halves of a loop = 9 loops." One Grade 8 student focussed on the relative size of the two rectangles. He wrote "the small one was 2/3 the size of the tall one, so 1/2 of 9 is 6." Although clearly representing only a small percentage of the students, this group appeared to have a good understanding of the comparison-by-multiplication aspect of proportional reasoning. Of the incorrect strategies, other than the"adders" discussed above, the next most common answer justification was pattern-based. For example, several students wrote "its a pattern" and then showed a two-by-two array with the numbers 6 and 8 placed beneath the numbers 4 and 6. Another student explained that "If there were two more the first time, then I add double the second time." A third example of a pattern-based response was given by a student who answered "4 loops because 4 + 6 = 10 for the first, so 6 + 4 = 10 for the second."

Item 12: Data analysis and summary information for this item are presented in Table 5. Comparatively speaking, a large percentage of students omitted this item. It is difficult to know if this was because students found the item difficult or they felt pressure to complete the test within a certain time limit. Of the two students who answered correctly, both indicated that the length of one metre of shadow would be about 2.6 (8+3), so the answer would be about 13 metres (5 x 2.6). One other student whose response I scored as incorrect explained that "8+3 is almost 3, and 14+5 is also almost 3, so the answer is 14." This response, and several others like it, clearly show multiplicatively-oriented conceptual understanding of the notion of proportion. Other examples include, "You need to double it and add some", and "For every metre you have "0.375 metres of a shadow, so you multiply 8 x 0.375."

Most incorrect responses reflected some form of "additive" thinking. The most common was for students to explain that they had to subtract 3 from 8 for the first flag pole and then add this difference of 5 onto the height of the second flag pole. This form of reasoning accounted for 14 of the 22 additive error responses of the Grade 7s and 8 of the 14 Grade 8s. Other additive error explanations involved adding and subtracting other combinations of the numbers 3, 5, and 8.

Table 4
Analysis of data with code summary Item 11

Code	Status/Explanation		Grade 7 (N=53)	Grade 8 (N=44)
2	Omitted		5 (9%)	9 (20%)
1	Correct		4 (8%)	6 (14%)
		each stick 1 1/2 loops	3	2
		half more per stick	0	1
		add 3 + 6	1	0
		small 2/3 of tall	0	1
		correct/no explanation	0	2
0	Incorrect		44 (83%)	29 (66%)
		add 2	30	27
		pattern 4,6,6,8	3	1
		pattern +2, +4	2	0
		pattern 4 to 6, 6 to 4	3	0
		explanation unclear	3	1
		no explanation	3	0

Table 5
Analysis of data with code summary for Item 12

Code	Status/Explanation	Grade 7 (N=53)	Grade 8 (N= 44)
2	Omitted	13 (24%)	17 (39%)
1	Correct	1 (2%)	1 (2%)
	one metre of shadow is 8+3	1	1
0	Incorrect	39 (74%)	26 (59%)
	errors with multi.ideas	2	5
	errors with add.ideas	22	14
	estimate or measured	2	0
	wrong answer/no explanation	7	6
	explanation not clear	6	1

Analysis by student

Considering the data on a student-by-student basis permits other forms of analysis (See Table 6). A pattern of CCCC means that this student answered all four items correctly. A CCNN pattern indicates that the student answered the first two correctly, but either answered incorrectly or omitted the second two. Working with the original coding sheet and the explanation categories list, a researcher could select and study a student typical of any one of the pattern groups, or could compare, for example, responses to the first two items of the CCCN pattern group with those of the CCNN group. I have chosen to focus on what appears to be three distinct groups of students: those with good proportional reasoning skills (CCCC, CCCN, and NCCC), those with limited skills (NNNN), and those perhaps in a transitional stage (CCNN, CNCN, CNNN, and NCNN).

There was a small group of students who demonstrated a good understanding of proportional reasoning. Of those in the CCCN category, three omitted the last question. The other three with errors on this question used multiplicative strategies: one is the student previously discussed who wrote that "8+3 and 14+5 were both almost 3, so the answer is 14." The second indicated 8+3 was 31/3; the third divided 3 by 8, rather than 8 by 3. The one student in the NCCC category is one of the students who divided 12 by 8 incorrectly, getting $1.40 instead of $1.50. Hence, in this group of good proportional thinkers, no student used an additive strategy. Each student correctly or incorrectly consistently applied multiplicative concepts.

Table 6
Frequency of correct and incorrect response patterns for Items 9 through 12.

Pattern	Grade 7	Grade 8
CCCC	0	1
CCCN*	2	4
NCCC	1	0
CCNN	17	13
CNCN	1	0
CNNN	9	8
NCNN	8	6
NNNN	15	12

*C: correctly answered; N: incorrectly answered or omitted

Of the 27 students in the NNNN category, one omitted all four items on the instrument and five others omitted three of the four. The majority of errors made by these students were either computational or reflected application of previously discussed inappropriate strategies.

The students in the third group appeared to be transitional in their understanding. With only one exception, students clustered around answering some combination of the first two items. These items did not require multiplicative strategies. Typical of this group is a student with an NCNN pattern, who for Item 9, correctly divided 12 by 8, multiplied the $1.50 answer by 2 but then multiplied that answer by 2 again to obtain $6.00 for his answer. For Item 10, he gave a correct halving/doubling reasoning strategy.

Discussion

From a review of the literature, it is clear that some authors advocate writing in prose in mathematics classes for both teaching and testing purposes. This study, which was designed to examine students' ability to respond in writing to ratio and proportion items on a conceptually oriented diagnostic test, would provide support for that view. Students explained their work with varying degrees of success. Some wrote terse remarks such as "I guessed." Others described in words the arithmetic operations that they had just completed. A number of students, rather than showing their work, just wrote the answers to questions but did provide appropriate explanations. There were also many students who responded with well-written sentences. Characteristic of these were the guessing-strategy answers like "First I guessed $1.00 but that was not enough to make the $12.00, so I tried $1.50...." These were the most interesting in that they provided clear indication of what the students said they did.

The writing technique (student written responses followed by coding, categorizing, and summarizing) deserves support as a viable research tool. A major concern with the writing technique, however, is that regardless of how well students described what they did, no student described why a particular strategy or algorithm was chosen. To explain the source of her answer of 72, a student wrote that she multiplied 6 x 12. She did not explain why she multiplied. Certainly, this is a major limitation to this form of data collection. But the writing technique does provide more information than traditional pencil-and-paper or multiple-choice tests. With multiple-choice tests, the researcher must infer what method the student used. With the writing technique, it is generally not difficult to determine what the student is doing. Still, think-aloud and interview approaches would be more beneficial in that students would not likely be able to omit questions and the researcher could ask for clarification of answers where needed. Further, these latter approaches would facilitate probing into why the student chose a particular algorithm or method.

Analysis of the results of this study does provide for interesting comparisons with those of Hart's (1981). Proportional reasoning has been the subject of much discussion. Among the chapters in the CSMS project report (Hart, 1981) was one devoted to ratio and proportion. In this chapter, Hart investigated aspects of ratios such as doubling and halving, finding rate per unit, and enlarging drawings in ratios of 2:1 and 5:3. She found that doubling and halving strategies were among the easiest for most children, enlarging non-rectilinear figures in the ratio 3:2 the most difficult. In the follow-up SESM project, Hart (1984) studied children's strategies and errors found to be common to a large sample of the scripts from the CSMS project. This second project involved diagnosis, analysis, and teaching by the researchers. Although findings confirmed much information obtained from the CSMS project, interviews and teaching allowed more in-depth probing of some misconceptions. For example, it was found that there was no evidence children used a standard ratio and proportion algorithm. They tended to devise their own methods. Post, Behr, and Lesh (1988) also discussed proportionality and focussed on methods likely to be used by students. They noted that proportional reasoning included notions of comparison, covariation, and the ability to process several pieces of information. This is perhaps the most important type of formal reasoning that students acquire during adolescence. Through their research they found that unit rate and factor strategies were the most successfully utilized.

In this study relatively few used the halving/doubling strategy. When given information about 8 tickets and asked about 6, almost 2/3 of the students opted for a unitary analysis procedure. However, for Item 10 involving the 2:3:5 ratio, more than half of the correct responses were obtained by halving the 2 then multiplying by the appropriate multiple. It appears that many students could use a halving/doubling strategy, but preferred a unitary analysis approach. The choice not to use the "intermediary 4" question did seem to influence the method chosen by the students.

Comparisons between the rectangles item and Hart's Mr. Tall/Mr. Short question show that "adders" made up a large portion of both populations. It is not clear that the use of the additive strategy resulted from the influence of the fraction or from some other notions inherent in the question. The concept of enlargement, or similarity, requires further study. With the last item on the test, it is difficult to assess the effect of the 3:5 ratio. A few students' explanations were clear and hence helpful but to examine the "fraction versus conceptual-understanding-of-proportion concepts" this item would have to be matched with comparable items with ratios of 3:6 or 5:10.

The above findings lend support to Hart's identification of levels of understanding of proportional reasoning where students progress from unitary analysis and factor/multiple methods up to more formal ratio approaches. Further support can also be provided in that, similar to Hart's research, no student in this study used the proportional statement of a:b = c:d. The teachers reported that their instruction emphasized informal ratio and proportion experiences, but that occasionally they did use this form. Hart found that less than 1 percent of her students used the proportional form.

Is this writing technique a useful strategy? A qualified "Yes" is perhaps the fairest answer. As a strategy it cannot replace interview approaches but it does appear to offer some advantages over traditional pencil-and-paper tests. In this study, the technique did permit some insights into students' conceptual understanding of proportional reasoning. It illustrated how the test could have been used to place students into groups with good, transitional, or weak ratio and proportion abilities. It also showed how the test might be used to identify various levels of understanding proportional thinking.

Areas for further research would include investigating whether or not tests could be designed that would study just one aspect of ratio and proportion in some detail; for example, having eight to ten questions on just enlargements. These tests would utilize writing or explanation-oriented approaches. Other studies could focus on whether or not the writing technique influences the students selection of strategy. Is memory decay affected? Might not the writing clarify some concepts? One other area of research that might be considered is related to the format of the test. Would multiple-choice tests combined with requests for children to explain their answers result in different responses from those obtained from the open response format? Would students respond differently if their response was not one of the options?

References

Ashlock, R. B. (1984) Focus on testing in diagnostic and prescriptive mathematics. *Focus on Learning Problems in Mathematics*, 9 (4), 49-54.

Bright, G. (1988) Story editing as a way to identify understanding of geometry. Paper presented at the annual meeting of the Research Council for Diagnostic and Prescriptive Mathematics, New Orleans, Louisiana.

Burton, G. (1985) Writing as a way of knowing in a mathematics education class. *Arithmetic Teacher*, 33 (4), 40-45.

Gordon, J. T. (1988) Writing: A teaching strategy for elementary algebraic fractions. *Focus on Learning Problems in Mathematics*, 10 (1), 29-36.

Hart, K. M. (Ed.). (1981) *Children's understanding of mathematics:* 11-16. London, England: John Murray (Publishers) Limited.

Hart, K. M. (1984) *Ratio: Children's strategies and errors*. Windsor, England: NFER-Nelson Publishing Company Limited.

Olson, J. (1986, April) Ready or not? Assessing developmental readiness for understanding algebra. Paper presented at the annual meeting of the Research Council for Diagnostic and Prescriptive Mathematics, College Park, Maryland.

Post, T., Behr, M. J. and Lesh, R. (1988) Proportionality and the development of prealgebra understandings. In A.F. Coxford and A.P. Shulte (Eds.), *The ideas of algebra, K-12*, (pp. 78-90). Reston, Va. : National Council of Teachers of Mathematics.

Stempien, M. and Borasi, R. (1985). Student's writing in mathematics: Some ideas and experiences. *For the Learning of Mathematics*, 5 (3), 14 17.

ASSESSING PROBLEM SOLVING IN SMALL GROUPS

Derek D. Foxman · Lynn S. Joffe

The work described in this paper took place in the context of national monitoring surveys of 11 - and 15 - year old pupils carried out in 1987 in schools in England, Wales, and Northern Ireland. The surveys were conducted under the auspices of the Assessment of Performance Unit (APU) at the Department of Education & Science (DES). About 800 pupils participated in the problem solving at each age level in groups of three. These pupils were sampled separately from the main samples of over 10,000 at each age level who took other assessments. Small-group problem solving was included among the assessments because of the growing educational interest in cooperative learning and in order to gain more information on children's performance in this area than in previous APU surveys (DES 1985). It was also expected that the problem solving processes would be more "visible" if externalised in discussion.

The development sought to devise a situation in which cooperative problem solving was likely to occur. A framework for assessment was constructed to enable the performance of groups of pupils to be rated by trained assessors on significant aspects of problem solving. In addition the assessors had available a scheme to categorise the group's activities and wrote detailed observations on the progress of the groups in their attempts to find a solution to each problem. As in other APU work, teachers played a prominent part in the development work and, during the surveys, as assessors.

Factors Facilitating Cooperative Learning

To what extent is cooperative learning in groups a feature of classrooms in Britain? For some years children in a high proportion of British primary classrooms have been organised in groups of 4, 5 or 6 around the tables in a room. The grouping is often by ability for mathematics or reading, and may be mixed ability or friendship groups for other areas of the curriculum (Great Britain, 1978). However, several research studies have shown that there is a distinction to be made between "grouping" and "groupwork" (Tann, 1981). Rarely do the classroom groups actually engage in collaborative work, nor are they asked to do so. More often they work as individuals; and, although neighbouring children may engage in discussion, this is not necessarily task-oriented (Galton et al., 1980).

In secondary schools grouping is rare, and groupwork even more so. However Cowie & Ruddock (1988), who have conducted a groupwork project in secondary schools, point out that the new 16+ examination in Britain, the General Certificate of Secondary Education (GCSE), is encouraging schools to provide more opportunities for cooperative learning. They found that nearly half of the new syllabuses make reference to groupwork in course aims and 20 percent in both aims and assessment objectives.

The potential benefits to children of learning in groups rather than individually has been of considerable interest to educationalists for some time. This interest stems from various sources: the desire to improve students' motivation, to develop their social and personal skills, or the need to organise learning with scarce classroom resources such as microcomputers. Theoretically, the work of Piaget (1959) and Vigotsky (1978) suggests that interactive situations should provide children with more opportunities for progression in their learning and development than working individually. Whether this can be demonstrated empirically has been the subject of a number of research studies in the past decade. For the purpose of this project it was important to know what factors were likely to facilitate cooperative working.

Slavin (1983) concluded that there is improved achievement in cooperative learning, but only when the group as a whole is rewarded rather than its members for their individual contributions. The type of task also influences the effectiveness of group situations (Cotton & Cook, 1982). Other factors which could affect group processes and achievement are the ability, racial, and gender mix within the groups. Higher attaining students could be expected to provide more explanations and more correct solutions to problems and may have more social influence within the group (Cohen, 1984). However, Webb (1982) reported inconsistent results from a number of studies of groups with similar or mixed attainment composition. Several researchers have noted differences in the behaviour of boys and girls within mixed groups (e.g. Lindow et al. 1985).

If cooperative learning does have more positive effects than individualised learning, to what factors can they be attributed? Piaget's (1959) view was that interaction with a peer pushes a child in the pre-operational stage towards considering more than one perspective

on a situation and so into the more advanced concrete operational stage (Mugny and Doise, 1978). Vigotsky (1978) considered that social interaction generally is a prime cause of intellectual development. Learning creates the "zone of proximal development" which is the distance between what children can do on their own and what they can do under adult guidance or in collaboration with more capable peers (Vigotsky, 1978). Light and Glachan (1985) found that children performed better when working cooperatively on a goal directed problem-solving task ("Tower of Hanoi") than when working individually, even when there was little verbal interaction. However, a task which produced more discussion ("Mastermind") was even more effective. They found that pairs of children at both 8 years and 12 years produced solutions in fewer moves than children working alone. Furthermore, groups who discussed the problem or argued about its solution most were significantly more likely to produce their solutions in fewer moves than those who argued least. Fletcher (1985) also found that groups were superior to individuals working on a microcomputer problem task, and that verbalising was a facilitatory factor.

Barnes and Todd (1977) conducted a study of the talk of 13 year old boys and girls while they were working on tasks set by their teacher. They reported that "the quality of discussion typically far exceeded the calibre of their contribution in class...". Barnes and Todd pointed out that talk in a classroom is usually managed by the teacher. In a group, without an adult present, it is the children who have to negotiate the course of the discussion with its episodes of silence and conflict and the need to encourage others rather than to dominate them.

These research studies do not present conclusive findings about the factors which might facilitate cooperative working in groups. The gender and ability mix and the extent of interactive talk within the group obviously needed to be considered. Another factor could be the gender of the assessor (Joffe and Foxman, 1988). The size of the group was not particularly noted as influencing cooperation in any of these studies, but it was an important factor in the organisation of the survey and so needed to be investigated in the development work.

Developing the Group Situation for the Surveys

The tasks tried out during the development, with the help of teachers' groups, were problems which could be tackled in different ways and had possibilties for extension. They included "everyday" tasks which required planning, and more purely mathematical problems which gave opportunities for pupils to conjecture relationships and test out their conjectures. Tasks were sought which, ideally, could be attempted by both age groups so that some comparisons might be possible between them. As previous research had suggested, the nature and quantity of the verbal interaction varied with the task. A lot of problems were tried and rejected: some produced a lot of animated talk, but little mathematics; others some mathematics but little discussion. Finally, four tasks were developed for both age groups with differences in detail between the two versions in each case. A fifth task was also developed for the 15 year olds. The tasks were:

Number Chains. Investigating the effect of applying a transformation rule to a number and then to the result of the transformation and so on successively, thus forming a chain of numbers. The rule used resulted in chains ultimately going into one of two repeating loops. The substantive problem was to find out what kinds of numbers led to a particular loop.

Filling trays. This was a version of the maxibox problem — finding the largest capacity of an open box or tray which results from cutting squares from the corners of a rectangular sheet of given size.

Class Trip/Day Out. Planning a day out on a limited budget given a map of places to visit, times of trains, activities and their cost, and menus at cafes.

Packaging. Designing a package to send three delicate glass spheres through the post.

Total 87 (for 15 year olds only). Devising a winning strategy in a game for two players or teams. Each team selects a number from 1 to 7 alternately, and the choices of both sides are added together. The first team to reach 87 is the winner.

The Class Trip (Age 11) and Day Out (Age 15) problems were borrowed directly from topics used previously in the 1-to-1 APU practical surveys. These topics were also used again in the group and individual test situations. The Number Chains and Packaging group topics were also adapted for the 1987 1-to-1 surveys, for comparison purposes, and a version of the Number Chains problem was adapted for a written test in the 1987 surveys.

Presenting the Tasks to the Groups

It was necessary to familiarise pupils with the content of the problem and what was required of them. Each session was divided into three phases. Phases I and III were interactive, while in Phase II the pupils were on their own, no help was allowed. In Phase I an introductory task was given which was related to, or part of, the substantive problem and prompts could be given. The main purpose of Phase I was to make sure that as many pupils as possible understood what they were asked to do in Phase II. In Phase II no interaction was allowed because it was found that, when it was, assessors became

part of the group and it was then difficult to get the group going on their own. Eye contact was avoided and, if pupils asked questions or asked for directions, they were given a neutral response: "That's up to you to decide." If a group attempted to draw in the assessor, the technique used was to feign a lack of interest. But when the group decided they had gone as far as they felt able to do, Phase III began in which clarification was asked for of what had not been clear to the assessor, and a general account given by the pupils of what they had done, why they had done it, and how they felt about the session. A fairly flexible script was used by assessors for Phases I and III.

During the session assessors wrote notes on what was happening and what was said in as much detail as they could manage. During the surveys a number of assessors exhibited great feats of concentration and the ability to record considerable detail (They completed their records after each session.). Tape recorders were not used, except in Phase III, because it was not possible to gauge their effect on the pupils before the surveys took place. Barnes and Todd (1977) in their research felt it was unreasonable to combine the effects of being tape recorded for the first time with that of working in groups for the first time.

The sessions lasted anywhere from 10-15 minutes up to an hour and a half. Most of the shorter and the longer sessions involved the more mathematical topics which, potentially, could be solved quite quickly. Some groups who could not find a solution were extraordinarily persistent in following up a range of hypotheses.

The size and composition of the groups

Decisions about the size and composition of the groups to be used in the survey were taken after a good deal of piloting to determine what kind of groups seemed to work best together. Usually only one piece of apparatus (e.g. a calculator) was provided in a group so as to emphasise the common aim, but boys were not infrequently observed to grab it. Girls could be at a particular disadvantage in such a situation, especially those from some ethnic minority backgrounds. Friendship groups were considered but it was found that, if one member was of a much higher attainment than the others, that person would be likely to dominate the group. Groups with more than 3 children tended to split into subgroups; it was more difficult for them to organise themselves and use the available resources effectively. Groups of only 2 pupils provided less discussion than larger groups. For these reasons it was decided to use groups of three pupils of the same sex and approximately the same attainment.

Assessing the Groups

The assessment schedule was developed with the assistance of groups of teachers experienced in using problem solving and investigative work in their classrooms, and guided by the work of theorists and researchers.

Theoretical perspectives on problem solving place stress either on the abilities needed for problem solving (Piaget, 1959; Krutetskii, 1976) or on the activities engaged in during the solution process (Polya, 1957; Schoenfeld, 1983). The APU is concerned with assessing performance and not with traits or capacities of the person and so ability models are of less interest in this context. Schoenfeld (1983) has given a detailed list of knowledge and behaviour necessary for what he believes to be an adequate characterisation of mathematical problem solving performance. The main categories are: Resources (e.g. facts, algorithmic procedures); Heuristics (e.g. drawing figures, introducing suitable notation); Control (e.g. planning, monitoring, decision making); Belief Systems - (One's "mathematical world view", determinants of an individual's behaviour).

In 1986 a number of marking schemes used for assessing investigative work in schools were collected and reviewed by NFER researchers. Most of them had been produced by teachers who had many years of experience of this activity in their classrooms. They were concerned with carefully written up *post hoc* reports of extended investigations and gave some useful indications of possible frameworks. The process objectives most common to the schemes were found to be largely compatible with Schoenfeld's ideas: Formulating the problem (Control); Use of mathematical strategies (Heuristics); Level of mathematics used (Resources); Evaluation and interpretation of results (Belief Systems). Because teachers are, in addition, interested in the way results are communicated and in an individual's personal contribution if the report is by a group, relevant categories covering these areas were noted in the review.

Normally the categories were marked on a scale of 4 or 5 points. Each point of the scale might carry an extended but fairly abstract description. For example a category, "Report as a communication", in one scheme describes a top-rated report as one which is: "Logically structured with suitable selection of what to present. Full explanation of the problem, its development and conclusions. Well written and appropriately illustrated with examples, tables and diagrams." A bottom-rated report would be "An untidy collection of results, badly organised, with little or no explanation."

Such descriptions must be relative to the normal standards of the material produced for assessment. Indeed, some teachers preferred to leave it there and simply stated: Marks 0 to 4 decided by experience of requisite standards.

Categories of performance do not constitute a model of problem solving and there was interest in the APU research in gathering data which might enable some general picture of the problem-solving process to be derived. Problem solving might be characterised as a cyclic activity which successively refines the direction which is taken towards a solution: for example, by formulating the problem more precisely; using more efficient methods etc. More realistically, it is likely to be untidy and opportunistic (Hayes-Roth and Hayes-Roth, 1979). Data from the sessions which might enable a generalised picture of the process to emerge would have to be detailed and chronological. There was interest too in determining relationships between various categories of performance and with background factors, such as the gender and ability of the groups.

In order to achieve these purposes, three sets of data were collected on the performance of each of the groups. In each case it was the group which was rated or categorised, not the individuals within it. It was made clear to the pupils at the beginning of a session that they were to work as a team and come up with an agreed solution. The data sets were: rating scales, a summary of performance, and observations.

Rating Scales. A number of scales were derived from the development work with teachers and guided by the theoretical and empirical work on problem-solving processes. There were eight main scales with sub-scales in most cases. They related to the areas of social interaction, problem solving skills, communication, and attitudes. Each scale or sub-scale had four points: 0 (low), to 3 (high). The scales were as follows:

1. **Social Interaction.** There was one sub-scale relating to the amount of cooperation and a set of categories defining the type of group.
2. **Awareness of Problem.** This category related to a group's overall grasp of what needed to be done to solve the problem: i.e. their overall strategy. Two sub-scales.
3. **Working on the Task.** The tactics used in relation to methods and level of mathematical argument. Three sub-scales.
4. **Resolution of the Problem.** This was an overall judgement of the group's performance, by the assessor on one sub-scale, and by the pupils of themselves on another, of the extent to which the problem had been satisfactorily resolved.
5. **Extension to Problem.** Very few pupils suggested additional questions which arose out of what they had done. Consequently, this category was largely redundant.
6. **Communication within Group.** There were separate sub-scales relating to oral, visual, and written means of communication.
7. **Communication with Assessor.** The three sub-scales related to the way in which a a group's report was presented in Phase III.
8. **Attitudes.** The three sub-scales related to the ratings of the pupils' involvement, persistence, and enjoyment.

Each point on each scale was described, in general terms for the age 11 survey. For the older pupils the descriptions of points on some of the scales were related more specifically to individual tasks.

Clearly there could be changes in the way groups operated during a session, and assessors were instructed that, in such cases, it was the later rather than the earlier behaviour which should determine the rating given. Thus a group which began cooperatively but ultimately worked individually should be given a low rating for cooperation, while one which began in a fragmented way but finally "gelled" should be given a higher rating. Similarly, in relation to Awareness of Problem, a group which began with little idea how to deal with a problem, but ultimately developed a good strategy, should be given a high rating. Not all of the sub-scales were relevant to every problem, and assessors were instructed not to give a rating if they thought a scale was inappropriate.

Summary of Performance. A second set of data was obtained from the assessors who were asked to summarise each group's performance under a number of headings. For example, the headings for Filling Trays were: Methods for finding the capacity of the trays; Accuracy of the methods used; Size of trays constructed; Accuracy of construction; Hypotheses generated about the relationships between the dimensions of the trays. For Class Trip the headings included Awareness of time in planning; Strategies and methods used; Awareness of cost; Recording.

Under each heading were listed the main possibilities which had been noted during the development work. The categories required assessors to make either yes/no decisions (Did the group find the capacity of trays by measuring, by using a cube, by multiplying length by breadth by height, by using a calculator, etc.) or ratings (Were the measurements made very accurate, not very accurate, or inaccurate?).

Observations of Group Activity. Observations were recorded by the individual assessors on A4 paper divided lengthwise into sections. One section was for the main observations. Other sections were reserved for comments on the group interaction, the processes being used by the group, and for recording the time at various points during the session. Assessors recorded in as much detail as they could during a session and then made up their notes when it was completed.

The Surveys

The assessors' task was to administer the assessments in the schools selected for the survey. Over a seven-day period in May 1987, for those participating in the age 11 survey, or November 1987 for those in the age 15 survey, they travelled to schools in England, or Wales, or Northern Ireland. At each school there were usually three groups of three pupils of the target age group and composition who were to be administered one problem each.

The assessors were experienced teachers nominated by their Local Education Authorities (LEA's) at the invitation of the NFER. A job description was sent to each invited LEA which emphasised that the teachers nominated should have taught boys and girls of the target age group and should be aware of recent developments in mathematics education. The most typical nominees for the age 11 survey were heads or deputy heads of primary schools or advisory teachers working across the LEA but with recent successful practice in the classroom. For the age 15 survey the nominees were heads of mathematics departments or advisory teachers. The locations of those invited were distributed as evenly as possible over the geographical area involved. There were 16 assessors altogether at each age level.

In previous APU surveys the majority of nominees for the practical tests were men. Spender (1981) has illustrated that mathematics teachers may respond differently to boys and girls, and so it was decided to control for any effects of gender of assessor in the 1987 practical surveys. LEA's were therefore asked to nominate an assessor of a specified gender. While an equitable gender balance of assessors was achieved for the age 11 survey, there was a slight imbalance in favour of men for the older pupils.

The main training provided was a two-day residential conference for each set of assessors held a few weeks before the respective surveys in May and November. Assessors were also expected to practice administering the assessments in their own schools between their briefing and the actual survey.

At the residential briefing the assessors were given the topic scripts and shown videotapes of groups working on the survey problems. There were sessions in which the teachers practised recording observations in detail, both from videotapes and with children from local schools. At the briefing for the secondary survey, groups of assessors also simulated the assessment situation: a technique which had been used successfully for several years at the briefing of the assessors of the APU 1-to-1 practical tests.

Some time was spent in discussing the nature of performance at different points on the rating scales. However, it was clear that a good deal more time was required than was available for the assessors both to observe and to reflect upon the wide range of ways in which pupils tackled the problems that had been revealed during the development work. Consequently, the way assessors interpreted the scales will be examined in the analysis.

The Design of the Survey

The number of schools participating was 100 in the primary survey, and 80 in the secondary. The schools were selected randomly in a stratified sample. Three pupils in each sample school were then selected randomly from among those in the target age range. Each of these pupils was assigned to one of the groups to be assessed. The final selection stage of making up the groups of three members was left to the school. The instructions from the NFER were for schools to choose two further pupils for each group, of the same sex and similar attainment to the pupils already selected randomly. While only one instance occurred where a school was unable to match the gender of a randomly selected pupil, there were a few cases where a very close attainment match was not possible.

There were two checks on the attainment mix within groups: schools were asked to give estimates of survey pupils' attainment within 20 per cent bands, and an independent estimate of attainment was obtained from the results of a written test taken by the same sample pupils. There was a different test for each age level but with similar content. The two tests were made up from the banks of APU written test items. The items selected were those relevant to the context of the problem tasks: measuring and spatial concepts relating to the Packaging task; reading tables and money calculations to the Class Trip and Day Out topics; number patterns to Number Chains; and area and volume questions to the Filling Boxes problem.

For the survey administration, topics were randomised over school and over assessors with the proviso that in every school the three groups took different topics. Thus there was no possibility that groups could glean any details of the problem they would be asked to solve from those pupils who had already been assessed.

This design resulted in about 70 groups taking each problem in the primary school and about 60 in the secondary. About 30 of the older groups took the fifth topic, Total 87.

Some Initial Results of the Age 11 Survey

The analysis of the results of the surveys of the two age groups are on-going; but, so far, only details of some the of age 11 results are available. These relate to the

ratings and provide indications of relationships between the scales and differences in responses to the four tasks. The importance of looking at the assessors' interpretation of these scales was stressed earlier. There are two ways in which this can be tackled: factor analysing the scales to examine their dimensionality, and relating the assessors' ratings to their detailed observations. The latter have not yet been analysed extensively, but some investigations of dimensionality have taken place. Factor analyses for each topic produced two main factors which were similar for all four topics. These could be described as cognitive and attitudinal factors. The cognitive factor had high loadings on the scales Awareness of Problem, Working on the Task, and Resolution of the Problem. The attitudinal factor had high loadings on the Amount of Cooperation and Attitude scales.

The following results are examples from one of the scales with a high loading on the cognitive factor — Resolution of the Problem (Assessors' Evaluation), and from one with a high loading on the attitudinal factor — Social Interaction.

Table 1
Ratings of Amount of cooperation within Groups

	Percent of groups rated as:			
Topic	0	1	2	3
Number Chains	2	10	42	46
Filling Trays	7	9	43	41
Class Trip	1	8	36	54
Packaging	7	15	56	21

The first question addressed concerns the extent to which the age 11 survey was successful in its aim of producing cooperative problem solving. The assessors' written comments and their discussion at a debriefing meeting held after the survey indicated that this was the case. This was reflected in their ratings of the amounts of cooperation for the four topics on a four point scale ranging from 0 (low) to 3 (high). These results are summarised in Table 1.

The two "everyday" topics, Class Trip and Packaging, received respectively the highest and lowest number of ratings of 3 for cooperation. This was almost certainly due to different task requirements: there was pressure in Class Trip to make decisions together, while some groups made individual designs for Packaging, although most reached a common final decision. It was encouraging to note the high cooperation ratings for the most purely mathematical topic, Number Chains.

Further information is provided by the assessors' categorisation of type of interaction within the group. Four main types of group interaction had been identified during the development. They could be placed on a scale of dominance of the group leader, from leaderless to authoritarian. Groups taking Number Chains were most likely to be non-authoritarian; again this is likely to be more a function of the task than of those groups who took the topic. Of the four topics it is the one where opinion, in contrast to logical argument, has least validity. The Packaging task had most scope for decisions to be made on the basis of opinion and therefore to be made by those who wanted to dominate. Girls' groups were given a much higher proportion of the top ratings for cooperation in 3 of the 4 tasks, but boys' groups had more of the top ratings in Number Chains. More girls' groups were classified as leaderless or were chaired in two tasks, the other two being more equable between the sexes in this respect.

Table 2
Ratings of Type of Group

	Percent of groups rated as:				
Topic	Leaderless	Chaired Group	Dominant Leader	Authoritarian Leader	Other
Number Chains	55	21	13	2	9
Filling Trays	49	23	12	3	13
Class Trip	44	23	18	4	11
Packaging	34	17	27	6	16

The assessor's evaluation was an overall summary rating of the extent to which a group resolved or solved a problem. Table 3 contains the distribution of ratings which were given by the assessors (0 low, 3 high):

Table 3
Assessor's Evaluation

	Percent of groups rated as:			
Topic	0	1	2	3
Number Chains	12	30	37	21
Filling Trays	7	42	31	20
Class Trip	1	14	61	24
Packaging	3	29	63	6

The "everyday" tasks appear to have been easier than the more obviously mathematical problems, although assessors were reluctant to give a top rating to the design task, Packaging.

Pupils selected for the group survey were due to take the special written test described earlier. Their score on this test gave an indication of the extent to which the pupils in a group had been of similar attain-

ment as requested. The test scores also provided a comparison of the small group sample with the main sample, some of whom had taken the same questions that appeared in the special test.

The result showed that the mean success rate of the questions in the small group sample test was significantly higher than that of the same questions taken by the main sample (51.3% to 48.0%). This finding is not all that surprising since two of the pupils in each group had been selected by the school and not randomly. So far as ability mix was concerned about two thirds of the groups had test scores within a range of 15 percentage points. There was an occasional extreme mix (e.g. 8%, 10%, 49% success).

Each of the test scores and the mean test score for a group was correlated with the assessor's evaluation and cooperation ratings for a group. This is summarised in Table 4.

Table 4
Correlations of Groups' Mean Test Score with Assessors' Ratings

Topic	Cooperation	Evaluation
Number Chains	0.50	0.64
Filling Trays	0.10	0.44
Class Trip	0.41	0.48
Packaging	0.21	0.33

Test score was not expected to correlate highly with amount of cooperation so it is interesting to note the relatively higher correlation for Number Chains, while cooperation was not associated with attainment for Packaging and Filling Trays. It should be noted that the assessors had no knowledge of the pupils' test scores (neither did their schools) nor were they informed of the school's estimates of pupil ability.

Conclusion

The results of the age 11 survey suggest that high rates of cooperative problem solving were achieved by groups of three pupils of the same gender and, mostly similar attainment. The amount and type of interaction was task dependent, the most cooperation being observed in groups taking the non-practical mathematical task.

The points on each rating scale were described in rather general terms for the assessors in the age 11 survey, but some were related more specifically to each topic for assessors of the older pupils. This may facilitate more differentiated rating between scales than was achieved at age 11 which resulted in two main factors only: cognitive and attitudinal.

Findings for the more detailed data which were obtained will be reported later.

References

Cohen, E. (1984). Talking and working together: Status, interaction and learning. In: P.L.Peterson, L.C. Wilkinson and M.Hallinan (Eds). *The Social Context of Instruction: Group Organisation and Group processes*. New York: Academic Press.

Cotton, J. and Cook, M. (1982). Meta-analyses and the effects of various systems. Some different conclusions from Johnson et al. *Psychol.Bull.* 92. 176-183.

Cowie C., and Ruddock, J. (1988) Testing Teams. *Times Educ.Supp.*15 April.

Fletcher, B. (1985). Group and individual learning of junior school children on a microcomputer-based task: social or cognitive facilitation? *Educ.Rev.* 37,3. 251-261.

Great Britain. Department of Education & Science (1978). *Primary Education in England*. London: HMSO.

Hayes-Roth, B. & Hayes-Roth, F.(1979). A cognitive model of planning. *Cognitive Science* 3. 275-310.

Joffe, L. and Foxman, D. (In press). *Communicating Mathematical Ideas*. London:HMSO.

Krutetskii,V.A. (1976). *The Psychology of Mathematical Abilities in Schoolchildren*. Chicago: University of Chicago Press.

Light, P. and Glachan, M. (1985). Facilitation of individual problem solving through peer interaction. *Educ.Psychol.* 5, 3 and 4. 217-225.

Lindow, J.A., Wilkinson, L.C. and Peterson, P.L. (1985). Antecedents and consequences of verbal disagreements during small-group learning. *J. Educ.Psychol.* 77,6. 658-667.

Mugny, G. and Doise, W. (1978). Sociocognitive conflict and structures of individual and collective performances. *European Soc.Psychol.* 8. 181-192.

Piaget, J. (1959). *Language and Thought of the Child*. 3rd Edition. London: Routledge & Kegan Paul.

Polya, G. (1957). *How to Solve It*. New York: Doubleday & Co.

Schoenfeld, A.H. (1983). Episodes and executive decisions in mathematical problem solving. In: R. Lesh and M. Landau (Eds.). *Acquisition of Mathematics Concepts & Processes*. New York: Academic Press.

Slavin, R.E. (1983). When does cooperative learning increase student achievement. *Psychol.Bull*. 94,3. 429-445.

Spender, D. (1981). *Invisible Women*. London: Writers and Readers

Tann, S. (1981). Grouping & Group Work. In: B.Simon and J. Willcocks. *Research and Practice in the Primary Classroom*. London: Routledge & Kegan Paul.

Vigotsky, L.S. (1978).. *Mind in Society*. Cambridge, Mass., Harvard Univ.Press.

Webb, N.M. Student interaction and learning in small groups. *Rev.Educ.Res*. 53,3. 421-445.

WIDENING THE PERSPECTIVE OF PROGRAM EVALUATION

David Nevo

Different educators mean different things when they use the word evaluation, and the evaluation literature provides multiple perceptions of evaluation. The well-known definition suggested by Ralph Tyler almost forty years ago, and still used by many, perceived evaluation as "the process of determining to what extent educational objectives are actually being realized."(Tyler, 1950, p. 69) This definition matched the general tendency in education to associate evaluation with testing and limited its scope to the measurement of students' achievements. Such an approach was also in congruence with the common sense of politicians and the general public who, on various occasions, requested that educators be accountable for their deeds and provide evidence of their effectiveness in the form of data on improvement in students' performance. Many evaluations of educational programs still focus on changes in students' achievement as a major variable for the assessment of the program. Even when some of them collect data related to the process of implementing the program being evaluated, it is used mainly as a means for interpreting the findings about students' performance, rather than as a criterion for assessing the quality of the program.

But evaluators experienced many problems in measuring the "really important" impacts of programs (e.g. long-range impacts). They also find it quite difficult to establish a causal relationship between students' participation in a new program and their achievement by means of implementing a true or even quasi-experimental design, as has been suggested by Campbell and Stanley (1966) and other research methodologists. Evaluators have also realized that the richness of a program or a project cannot be expressed only by its impact on students' behaviors, nor can the full range of their clients' information needs be served by data only on students' test scores.

The evaluation literature has been suggesting for some time many attempts to extend the scope of information that should be collected regarding each program that is being evaluated. Stake (1967) in his Countenance Evaluation Model suggested that two sets of information be collected regarding the program being evaluated: descriptive and judgmental. The descriptive set should focus on intents and observations regarding antecedents (prior conditions that may affect the outcomes of the program), transactions (the process of implementing the program), and outcomes of the program such as students' achievements but also other outcomes. The judgmental set of information in Stake's model is comprised of standards and judgments by relevant audiences regarding the same antecedents, transactions, and outcomes.

Guba and Lincoln (1981) extended Stake's approach and applied it to the naturalistic paradigm. They suggested that the evaluator collects five kinds of information as follows: 1) descriptive information regarding the program, its settings and its surrounding conditions; 2) information responsive to concerns of relevant audiences of the evaluation; 3) information about relevant issues; 4) information about values; and, 5) information about standards relevant to the worth and merit of the assessments.

Stufflebeam, together with a prominent group of evaluators (Stufflebeam, et al., 1971) analyzed various types of decisions and decision-making settings. They endorsed Stufflebeam's CIPP Evaluation Model, suggesting that evaluation focus on four sets of information regarding the program being evaluated: the goals of the program, its design or strategy, its process of implementation, and its outcomes.

The notion that a wide range of information should be collected regarding each educational program has been supported by many other authors in the evaluation literature published in recent years (e.g. Meckenzie, 1983; Nevo, 1983; Dorr-Bremme, 1985; Colley and Bickel, 1986; Glasman and Nevo, 1988). This was also the perspective of our evaluation study of an elementary school computer assisted instruction (CAI) mathematics program (the TOAM program). Nevertheless, in planning the evaluation study, we had to work hard to convince our clients that "hard data" on student achievement is not the only thing that could be useful to them in making decisions about the program. And since similar difficulties have also been experienced in other evaluations, we would like to reemphasize the importance of widening the perspective of program evaluation, to point out some possible methodological solutions, and to discuss the utility of such an approach on the basis of our experience.

The Program and its Evaluation Design

The TOAM program is an Israeli adaptation of a CAI mathematics program initially developed at Stanford University in the early sixties by Suppes and associates (Suppes, et al., 1968). The program was adapted to the local mathematics curriculum and has been used in Grades 2 to 6. Participating students used the computer twice a week, each time for 20 minutes, where they had an opportunity to practice individually

on a graded sequence of exercises and were provided with feedback regarding their performance. The teacher is also provided with a diagnostic summary of the whole class at the end of each period. The computer in this program is used only during 40 minutes a week out of a total of about four hours of weekly mathematics instruction. The computer is used only for practice and diagnosis while most of the instruction is done within the regular class by means of other teaching methods.

The evaluation was conducted within the framework of the city of Tel Aviv, where the local department of education decided to introduce the TOAM program into schools with a high proportion of culturally disadvantaged students. The purpose of using the program was to help low-achieving students without hindering the progress of advanced students. TOAM computers had been used for some years in the schools of Tel Aviv when the local department of education decided to fund a one-year evaluation to examine the usefulness of the program and how it could be improved.

In light of our perception regarding the scope of evaluation (Nevo, 1983), and on the basis of interaction with our clients and other stakeholders in the program, three major questions were identified as reflecting what might be their main information needs. The following evaluation questions were agreed upon to be addressed by the evaluation:

a. Are the rationale and the structure of the TOAM program based on acceptable educational approaches providing a reaonable chance to affect its target population?
b. Is the program being implemented as planned and in an efficient way?
c. Does the TOAM program have an impact on students' achievement in mathematics and on their attitudes towards studying mathematics?

Four sources of information were used to address the first question. They were: major documents of the program; interviews with the program personnel; a review of the literature on mathematics education and on computer assisted instruction, including some meta-analysis studies; and experts' opinions on the program, obtained from four experts especially for this evaluation.

The second evaluation question was addressed by means of: administrative reports of the program; structured observations in mathematics classess and computer practice sessions (46 observation hours in 9 schools); interviews with teachers and computer personnel; and questionnaires administered to students (n = 241), teachers (n = 191) and principals (n = 16).

For the third evaluation question data were collected on students' achievement and their attitudes toward mathematics. Data on TOAM computer scores were analyzed for a total of 5254 students in Grades 2 to 6 in 19 schools. Standardized paper-and-pencil tests were administered to 273 TOAM students in Grades 4 to 6 and to 214 students in comparison groups. Attitude questionnaires were administered to 123 TOAM 6th graders and to 118 students in similar comparison classes. Students in comparison groups were selected on the basis of similar socio-economic background to that of the TOAM group but random assignment of students to groups was not feasible in this study.

Major Findings of the Evaluation

A detailed presentation of the data analysis procedure and findings of this evaluation can be found elsewhere (Nevo, 1984; Metzer, 1986; Nevo, in press). In this paper only a summary of the major findings will be presented as a basis for our discussion on the scope of evaluation. Following are major findings regarding each evaluation question:

Are the rationale and the structure of the TOAM program based on acceptable educational approaches providing a reasonable chance to affect its target population?

a. TOAM is based on a behavioristic appoach emphasizing the relationships among stimulus, response, and reinforcement. This approach was highly criticized in the literature and by the experts used in this evaluation as an approach of limited value appropriate mainly for learning simple tasks.

b. An extensive review of the literature on CAI and mathematics education showed that the use of computers in instruction can be useful when used in conjuction with regular class instruction, and with close cooperation between the teacher and the computer.

c. Previous studies, conducted by the organization which developed and operated the TOAM program around the country, which showed the effectiveness of TOAM in improving students' achievements in mathematics were all based on TOAM computer tests rather than on standardized paper and pencil tests.

Is the program being implemented as planned and in an efficient way?

a. Review of administrative reports and direct

observations in schools showed that the operation of TOAM within the schools was well organized and implemented according to formal instructions and with almost no complaints from participating schools.

b. The bi-weekly computer practice sessions were implemented by special TOAM instructors; the participation of class teachers in those sessions was very limited. One third of the teachers indictated in their questionnaires that they do not attend regularly the computer practice sessions with their students. Our sample observations showed that in more than half of the the sessions teachers were absent. There were no regulations regarding teacher presence in computer practice sessions of their students.

c. More than 80 percent of the teachers indicated in their questionnaires that they used the computer reports, provided at the end of each practice session, in planning their lessons. However in our structured observations in 32 fourth and sixth grade lessons we succeeded in tracing some kind of reference to computer reports in only one third of the classes.

d. The teaching style of teachers in classes participating in the TOAM program was found (in classroom observations and teacher questionnaires) to be similar to the typical teaching style of teachers in regular classes in Israeli schools and included very little work in small groups and individual work of students. However, the tendency to use "non-conventional" teaching methods was slightly stronger among teachers who had participated in the program for more than one year.

e. Teachers seemed to be pleased with the orientation training that they got when they joined the TOAM Program, but many of them (30 to 50 percent) asked for additional guidance in teaching gifted students, working in small groups and dealing with individual differences in heterogeneous classes. More than 50 percent of the teachers did not get any in-service training during their first year in the program except a one day orientation when they joined the program.

Does the TOAM program have an impact on students' achievement and on their attitudes towards studying mathematics?

a. Analysis of TOAM computer test scores in participating schools showed that the percent of students reaching the expected minimal requirement level for their grade at the end of the year was significantly higher than an estimated level of non-participating students. However, a high percentage (33 to 85) of participating students in various classes did not reach the minimum requirements determined by the TOAM program by the end of the year .

b. The analysis of the TOAM computer test scores also showed that the progress of high level students was significantly greater than the progress of low level students. Thus, the gap between low achievers and high achievers seemed to increase by virtue of the TOAM program.

c. Standardized paper-and-pencil tests administered to 4th and 6th grade students participating in the program and to non-participating students with similar backgrounds showed no statistically significant difference between the overall mean scores of both groups. However, in two out of the six subscores of the fourth grade test, a significant difference in favor of the TOAM group was found. No significant differences in subscores were found in the sixth grade, but a significant difference was found among the groups in the percentage of students who got high scores on the entire tests (more than 75 percent correct answers).

d. Regarding students' attitudes towards mathematics, "math anxiety" was found to be significantly lower in the TOAM group in the sixth grade compared to the comparison group, but no significant differences among the groups were found regarding other sub-scales of the attitude questionnaire.

e. Teachers and principals expressed overall positive attitudes towards the program and thought that TOAM had a positive impact on students' achievement, especially on good students.

Summary and Discussion

In spite of the fact that during the planning phase of the evaluation our clients showed a strong preference for information on students' achievement, that would demonstrate the impact of the TOAM program, such information turned out not to be useful when the evaluation study was concluded. Since, as we mentioned earlier, the use of an experimental design within the framework of this study was not feasible, there were

some limitations on the inference that could be made from our data on students' achievement in the TOAM groups and the comparison groups. However, it seemed clear that there is no strong evidence to support the claim that TOAM has a significant impact on students' achievement, and that such a claim is unwarranted at least considering the way the program has been actually implemented. The contradictory findings for the computer tests and the paper-and-pencil tests were interesting. So were the findings that showed that TOAM, which was funded within the framework of special support to disadvantaged students, seemed to be increasing the gap between low achievers and high achievers.

But the important question was, what could be done with those findings regarding the impact of TOAM? Soon it became clear that the answer was: "Really not much!" Nobody would make a decision to discontinue the TOAM program in the Tel Aviv schools, since there was no available alternative on the market that could offer a complete set of courseware in mathematics for elementary school classes. It was also apparent that no one would shift funds from a CAI program to other educational projects at a time when the whole educational system seemed to be "hooked" on computers and perceived the introduction of computers into the school as a major effort to modernize the educational system. Actually, if one would be willing to decide to discontinue funding of the TOAM program he could do so on the ground of a simplistic rationale and poor implementation as was found in our evaluation.

When we submitted our final evaluation report it was apparent that although the original charge of the evaluation was formative as well as summative, its major contribution could be only in its formative mode. TOAM was there to stay, and the only decisions that could be made about it would be related to its improvement. But not much advice could be derived from the test results, at least not as much as could be derived from the other findings.

Our findings regarding the rationale of the program and its structure (first evaluation question) suggested clearly that TOAM was based on a simplistic approach that has been highly criticized by experts on CAI and mathematics education as well as by the research literature. Our study also showed (second evaluation question) that teachers were not getting sufficient training and guidance to incorporate the work of their students with the computer into the whole process of teaching and learning mathematics. On the basis of these findings it was quite simple to develop recommendations regarding the improvement of the rationale of the program, the structure of its courseware and its use in the school. Among other things we recommended that the organization developing and administering the TOAM program seek advice from the current literature and additional specialists in CAI and mathematics education to update its courseware and renew its conceptions. We also recommended that an extensive manual for TOAM teachers be developed and that an effective teacher training and guidance program be developed and implemented.

Obviously, we must continue to seek evidence on the impact of educational programs as part of our evaluation practice. But, it is also very important to include in our evaluations activities directed toward the assessment of the progrram rationale and its strategy and process of operation. If we decide to follow this advice, we will find that there are sufficient tools to do so; some of them old, and some of them quite new. In this regard we should remind ourselves of observational techniques (e.g. Simon and Boyer, 1976), content analysis methods, use of experts' opinions (e.g. Nevo, 1985), and the use of recently developed methods of meta-analysis (e.g. Hedges and Olkin, 1985) for quantitative synthesis of research literature.

References

Campbell, D.T. and Stanley, J.C. (1966). *Experimental and Quasi-Experimental Designs for Research* Chicago: Rand McNally.

Cooley, W.W. and Bickel, W.E. (1986). *Decision-Oriented Educational Research* Boston: Kluwer-Nijhoff.

Dorr-Bremme, D. (1985). Ethnographic Evaluation: *A Theory and Method. Educational Evaluation and Policy Analysis* 7(1).

Glasman, N.S. and Nevo, D. (1988). *Evaluation in Decision Making: The Case of School Administration.* Boston: Kluwer Academic Publishers.

Guba, E.G. and Lincoln Y.S. (1981). *Effective Evaluation.* San Francisco: Jossey-Bass.

Hedges, L. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Orlando, FL: Academic Press.

Mackenzie, D. E. (1983). Research for School Improvement: An Appraisal of Some Recent Trends. *Educational Researcher*, 12(4): 5-16.

Metzer, D. (1986). Evaluation of the Project of Teaching Mathematics with TOAM Computers. Unpublished M.A. Thesis, Tel Aviv University, School of Education (in Hebrew).

Nevo, D. (1983). The Conceptualization of

Educational Evaluation : An Analytical Review of the Literature. *Review of Educational Research*,53 (1): 117-128.

Nevo, D. (1984). Evaluation of the Tel Aviv Enrichment and Reinforcement Project: Final Report. Tel Aviv: Tel Aviv University, School of Education (in Hebrew).

Nevo, D. (1985). Experts' Opinion: A Powerful Evaluation Tool. Paper presented at the Annual Meeting of the American Educational Research Association. Chicago, March 31-April 4.

Nevo, D. (in Press). *Useful Evaluation*. Tel Aviv : Masada Press (in Hebrew).

Simon, A. and Boyer, E.G. (Eds.) (1976). *Mirrors for Behavior: An Anthology of Classroom Observation Instruments*. Philadelphia: RBS.

Stake, R.E. (1967). The Countenance of Educational Evaluation. *Teachers College Record* 68: 523-540.

Stufflebeam, D.L. et al. (1971). *Educational Evaluation and Decision Making* . Itasca, Ill.: Peacock.

Suppes, P. et al. (1968). *Computer Assisted Instruction: The 1956-66 Stanford Arithmetic Program*, New York: Academic Press.

Tyler, R. W. (1950). *Basic Principles of Curriculum and Instruction*. Chicago: University of Chicago Press.

ASSESSMENT OF OPEN-ENDED WORK IN THE SECONDARY SCHOOL

Dylan Williams

"I don't know what you what you mean by 'glory'," Alice said. Humpty Dumpty smiled contemptuously.

"Of course you don't - till I tell you. I meant 'there's a nice knock-down argument for you'!"

"But 'glory' doesn't mean 'a nice knock-down argument'," Alice objected.

"When I use a word," Humpty Dumpty said, in a rather scornful tone, "it means just what I choose it to mean - neither more nor less".

"The question is," said Alice, "whether you can make words mean so many different things."

"The question is," said Humpty Dumpty, "which is to be master - that's all." (Carroll, 1871)

What kind of activity?

There does not appear to be any broad consensus about the meaning of the terms "open-ended activity", "problem", and "investigation" when applied to school mathematics. For the purpose of this paper, therefore, I shall use "investigating" to describe the entire spectrum of mathematical activity. This ranges from becoming aware of a domain to be explored, through defining or posing a problem, solving the problem as defined, to extending or reformulating the problem, and then, possibly, going around the cycle again. "Problem solving" is then a distinct phase in "investigating", as, for example, is "problem posing". What precisely constitutes "mathematical activity" is, of course, principally a question about the nature of mathematical knowledge; in other words, of epistemology.

Many distinctions in the nature of knowledge have been proposed. Some of these are intended to apply principally to the domain of mathematics, while others are much more general See the list below for some examples.

Most of these distinctions appear to have some commonality; they seem to be addressing different aspects of the same kind of idea. Rather than invent a new pair of words I shall use the term "conceptual knowledge" as a generic term for the kind of knowledge typified by the entries of the first column of the above list and "procedural knowledge" for those in the second column.

The emphasis in much recent research, especially that done in North America, appears to have been on how procedural knowledge becomes transformed so that it also exists as conceptual knowledge. This can be interpreted as reflecting the concern of research to make the "traditional" teaching of mathematics more effective (By "traditional" I mean teaching where mathematical knowledge is "laid out" before the learner, and the learner "makes sense" of it). Furthermore, the main focus of this research has been "bottom up" in that it has concentrated on relatively simple (but still very complex!) domains such as young children's understanding of arithmetic. In contrast to this approach, it is possible to focus primarily on conceptual knowledge, and concentrate on how knowledge that exists initially as conceptual knowledge can become "routinised" or "made automatic" so that it also exists as procedural knowledge. Learning that takes place in this way highlights the distinction between procedural knowledge that is "backed up" by conceptual knowledge, and procedural knowledge that is mainly isolated in that domain. However, for a given activity that we might use for assessment, we cannot be sure that we are assessing conceptual rather than procedural knowledge. For example, solving a quadratic equation is, essentially, a test of procedural knowledge if you know the formula. If, on the other hand, you don't know that there exists such a formula, then the task is much more

	"conceptual"	"procedural"
Katona (1942)	meaningful apprehension of relations	senseless drill and arbitrary associations
Maier (1945)	productive thinking	re-productive thinking
Wertheimer (1959)	structural understanding	rote memory
Scheffler (1965)	knowing that	knowing how to
Tulving (1972)	semantic memory	episodic memory
Greeno (1973)	propositional knowledge	algorithmic knowledge
Skemp (1976)	relational understanding	instrumental understanding
Piaget (1978)	conceptual understanding	successful action
Anderson (1983)	declarative knowledge	procedural knowledge
Hiebert & Lefevre (1986)	conceptual knowledge	procedural knowledge

likely to test your conceptual knowledge. Such examples are not confined to the traditional school mathematics curriculum. For example if we have a cube, and each face is to be painted either black or white, how many distinct arrangements are there if rotations and reflections are not to be counted as different? This is certainly a non-routine task for most, but if you know the Polya-Burnside formula, it is just a matter of following the steps.

Here I would like to introduce the idea that certain mathematical tasks might act as "amplifiers" of the differences between different students' previous experience. If the items quoted above were given to 16-year old students, it seems likely that the quadratic equation item would tend to increase the effects due to differences in students' past experience, while the cube-colouring task would tend to reduce those effects. Whether there exist tasks that "reduce experience" sufficiently to be useful in this respect and whether such tasks can reliably be selected are issues for debate, but I feel that a focus on this kind of activity will provide us with insights into the general processes of mathematics. Mathematics has often been viewed in the past as "what is left when all the context has been removed." In this sense, I am proposing that by rendering useless as much of a student's procedural knowledge as we can, we can learn much about the essence of mathematical thinking.

Here, I want to make clear that I am not advocating that the procedural knowledge that a student has is not important. What I am arguing is that by using tasks that reduce its effect on student performance (and perhaps only then), we can begin to look at more general mathematical processes. However, these mathematical processes are useless without mathematical objects upon which to operate. Ultimately, therefore, I see the assessment of these processes as complementary to more traditional forms of assessment, rather than replacing them. To heighten the contrast with the existing paradigm further, this approach can be applied, not to relatively simple domains like arithmetic, but to relatively complex domains like students' attempts at solving complex mathematical problmes. This immediately raises two questions: how can we engender this kind of activity and how do we assess them?

What kinds of task?

The relationship between task and activity is clearly far from straightforward (See, for example, Christiansen and Walther, 1986). Bauersfeld (1979) has pointed to the differences that often exist between the matter intended, the matter taught, and the matter learned. Burton (1980), on the other hand characterises the important distinction as being between puzzles and problems. What is repeatedly stressed is the importance of the student making the task her own. Any attempt to understand when and how this happens cannot be based on an analysis of the task alone, or on just the cognitive and meta-cognitive characteristics of the student. This realisation is manifest in the notions of belief systems as used by Schoenfeld (1985), situational analysis (Balacheff, 1985; Depuis 1985), and perhaps most significantly in activity theory (Christiansen and Walther, 1986; Mellin-Olsen, 1987). It is in connection with these non-cognitive factors that the idea of an open-ended activity becomes important. Schoenfeld (1985) has reported (as have many others) that students' attempts at tasks are often distorted by their beliefs. If they think that the teacher has a particular answer in mind, the students will often not be thinking mathematically, but will, instead, be trying to "guess what's in teacher's head." I will therefore use the term "open-ended activity" for a task which presents a more or less clearly defined starting point for a student, but where the exact nature of the goal, and consequently when the activity terminates, is under the control of the student.

To summarise, the stance that I am adopting here is that there do exist tasks that generate activity in students that reduce the effects of procedural knowledge sufficiently to allow us to assume that that the degree of success on those tasks is primarily due to conceptual knowledge; they are valid in that the activity that they generate is, in essence, mathematical; they can be presented to students in such a way as to cause the students to "engage" and "make them their own." Here are some candidates:

> How many integral-sided triangles can be made with longest side "n"?
> How many integral-sided triangles can be made with perimeter "n"?
> How many ways are there of giving someone "n" cents?

What kind of assessment?

In the United Kingdom over the past few years, these "open-ended" tasks have been increasingly used in the teaching of mathematics. A broad consensus does seem to be emerging that any mathematics curriculum which neglects these aspects of the learning of mathematics is deficient in important respects (HMI, 1985). However, while there is much evidence of this kind of activity in classrooms, very little research has been done on how these kinds of thinking might be assessed or evaluated. The major approaches to assessing mathematical activity can be classified by the principal variable used to evaluate the quality of the thinking involved. In the "cognitive demand" approaches, the central feature is (adopting a metaphor from competitive diving) the "degree of difficulty" of the task; or, where there is a series of tasks, the hardest task succesfully attempted. If we persist with the diving metaphor, the other approaches can be thought of as assigning central status to the "marks for style." The most

important feature is the extent of progress made on a single task.

"Cognitive Demand" Approaches

In the literature from cognitive and developmental psychology, the work of Piaget (1956), Pascual-Leone (1970) and Case (1985), offer us a number of possible cognitive structures that might be used as the basis of an assessment scheme. The major drawbacks of these schemes are two-fold. In the first place, they tend to have a relatively small number of levels; and, secondly, they tend to be rather difficult to apply to the complex mathematical tasks under consideration here. This appears to be principally because the major research instrument used for assessing the level of development tends to be a graduated series of simple tasks rather than a single complex task.

In geometry, the model proposed by Van Hiele (1986) can be used to assign students' geometric thinking to one of five different levels. Here the emphasis has moved slightly away from cogntive structures and a little towards levels as existing in the organisation of the thinking of the individual. This idea is more completely realised in the SOLO taxonomy developed by Biggs and Collis (1982) which completely eschews the idea of a "hypothetical cognitive structure". SOLO is an acronym for the Structure Of Learning Outcomes; and, as its name suggests, it concentrates on assessing the quality of the learning outcome, without speculating about how it was achieved.

These last two models offer significantly more scope for the assessment of complex mathematical tasks, because they deal with complex tasks; however they share two drawbacks. The first is that the degree of resolution of the assessment instrument tends to be small. The Van Hiele scheme gives about three levels for the mathematical attainment of the age-16-cohort, and the SOLO taxonomy gives about five. This is, of course, a recurring theme; the more levels you get, the less reliable is the allocation of a given piece of work to a particular level. The second drawback is that these schemes do not appear to transfer in any simple way to the kinds of activity being discussed here. It seems, therefore that both the model offered by Case, and the SOLO taxonomy offer considerable scope for the future, but appear too difficult to translate into assessment practice at the moment.

"Extent of Progress" Approaches

Drawing on the work of Polya (1945), Schoenfeld in the United States, and Mason and Burton in the United Kingdom, have developed heuristic models of the problem-solving process (See, for example, Schoenfeld, 1985; Mason, Burton & Stacey, 1982; Mason, 1984; Burton, 1984). These heuristic-based schemes appear, in turn, to have informed the various schemes that have been developed in the UK for the large-scale assessment of mathematical problem-solving, investigation, and exploration. Examples of these are the Department of Education and Science's Working Party on Mathematics Draft Grade Criteria (SEC, 1985), the assessment model proposed by the Oxford Certificate of Eduational Achievement (OCEA, 1987a; 1987b), and all the assessment schemes proposed by the examination boards for the examination of coursework in GCSE. Other work in this same tradition of assessing mathematical process has centered on the work of Bell. In a series of studies (Bell 1976; 1979; Horton, 1979; Galbraith, 1981), Bell and others have examined students' proof-explanations and have elicited structures that are quite general.

All these process-based schemes have tended to regard the cognitive demand of the task as of secondary importance, and, in effect, therefore, treated all tasks as essentially equivalent. Consequently, these process-based schemes would not distinguish between the same process displayed in different problem-contexts, even though the difficulty (as determined by, say, facility) might be very different. Clearly then, what is required is a scheme that can combine the "cognitive demand" approach with the "degree of difficulty" approach. Such a scheme is probably a long way off, but what follows is an outline of a way in which account can be taken of the degree of difficulty of the task, so that the process-based schemes referred to above can be used with greater precision. As outlined above, the stance adopted in this paper is explicitly constructivist in the sense outlined by, for example, Davis (1984) and Novak (1986). All the students' actions are assumed to be "intelligent" within the frame of reference of the student. In assessing the activity we are seeking to locate that frame of reference, and as far as is possible, assess it on its own terms. No account is taken of the relationship between the task intended by the teacher, and the activity in which the student engages. All that is important is how difficult the "mathematical terrain" was to chart, and the quality of charting done.

Task variables

The tasks that have come to be most frequently associated with open-ended activity in the UK can be characterised as Data-Pattern-Generalisation (DPG) tasks (Wells, 1986, p11). Having defined a problem, the student typically generates some data, organises the data, looks for patterns, makes hypotheses, tests them, and, if possible, proves them. The three main phases of activity are therefore systematic generation of data, deriving relationships, and making proofs. In this paper I will deal only with the first two of these. For detailed accounts of students' proof-explanations see Bell (1976, 1979, 1980) and Galbraith (1981).

Systematic generation of data

In the task called "Sending cards" (GAIM, 1988), students are asked to investigate how the number of cards sent varies with the size of the group if everyone in the group sends a card to everyone else. Most students generate the data systematically here by incrementing the independent variable (the number of people in the group) by one each time, giving rise to the sequence 2, 6, 12, 20, 30, ... (i.e. twice the triangle numbers). In the task "How many rectangles?" (SMILE, 1975) students are asked to investigate how many rectangles are created if a number of horizontal and a number of vertical lines are drawn across a rectangle. This situation is clearly more complex than that in "Sending cards", in that there are two independent variables: the number of horizontal lines, and the number of vertical lines. Most students who manage to generate the data systematically do so by holding one of the independent variables fixed, and incrementing the other.

Unfortunately, characterising the complexity of a task by the number of independent variables breaks down when we consider a task like "Four squares" (GAIM, in press). Here students are required to generate all possible colourings of a four-region map with four colours, each colour being used exactly once. However, we can generalise the notion of the number of independent variables by introducing the notion of a search-space. The search space of a task consists of all possible combinations of the values of the independent variables. The difficulty of carrying out a search is then characterised by the efficiency of various search strategies in exhausting the space.

At this point it is probably worth noting that this idea of "search space" is different from the idea of a "state-space" in the problem solving literature. Searches of "state-spaces"are designed to reach one particular state (the goal state). In this case, the object is to locate every element of the search space. The strategy used above for "Sending cards" can be termed a linear search strategy, or a 1-dimensional Cartesian search strategy. In the same way, the strategy used for "How many rectangles" would be termed a 2-dimensional Cartesian search strategy. Using a 4-d Cartesian search strategy on "Four squares" will yield all the elements of the search space, but only at the expense of a considerable number of "disallowed" combinations. In fact it will yield 256 combinations of which only 24 are allowable a rejection rate of over 90 percent! However, we (and most students who attempt this activity) can do better than this by using a "tree-like" search strategy. This strategy generates only allowable combinations, and generates all the possible combinations without repeating any of them. It is, in fact, the most common strategy employed by students who are successful in finding all the combinations.

To sum up then, "Four squares" is exhausted by a 4-d Cartesian search strategy, but it is not efficient, while the tree-like search strategy is both efficient and exhausting. These strategies can also be thought of as similar to the production systems used in, for example, Anderson's ACT theory (Anderson, 1983). We can go on to consider tasks for which efficient strategies do not exist. A good example is the task of finding all the pentominoes, in other words finding all the ways of arranging five squares if all the squares must join edge to edge and corner to corner. Most recalcitrant of all are those search spaces for which there is neither an efficient nor an exhausting strategy.

Deriving relationships

Having derived the data, the next stage is to look for patterns within that data; and, where possible, to hypothesise relationships. Clearly if the value of the dependent variable is always one more than that of the independent variable (e.g. the relationship between the number of fences and fence-posts) that relationships is going to be easier for students to discover than in, for example "Sending cards". Another aspect of the complexity of the mathematical relationship between variables is the way that students choose to express the patterns that they discover. For example, in "Sending Card"', students seem to find it easier to describe the sequence as "going up in even numbers", than as "the number of people times by the number of people minus one". The first is an example of a term-to-term rule, while the second is a position-to-term rule. In general, the term-to-term rule is "easier" and so more accessible to students. This distinction has actually more to do with how students represent their discovery than with the structure of the problem, and properly belongs in the heuristic- or process-based side. However, I have mentioned it here, because there are situations where there is no position-to-term rule, but there is a term-to-term rule (See, for example, the Josephus problem in Engel, 1985, p185). The following list is offered as a tentative hierarchy. It is not particularly "robust" since very large numbers, for, say, an additive mapping might be harder than small numbers with a linear relationship: additive, multiplicative, linear, quadratic, polynomial, exponential, other (e.g. involving hcf or gcd).

Summary

This paper has presented a model for evaluating the "degree of difficulty" of a class of mathematical activities which can be used to complement heuristic- or process-based assessment schemes in order to give a more accurate indication of the "power" of the mathematical thinking represented by a piece of work. The model characterises this "degree of difficulty" by two factors: the structure of the search space of the problem, and the complexity of the mathematical relationship between the variables.

References

Anderson, J.R. (1983). *The Architecture of Cognition*. Cambridge,MA:Harvard University.

Balacheff, N. (1985). Cognitive versus situational analysis of problem-solving behaviours. *For the Learning of Mathematics* 6 (3) 10-12.

Bauersfeld, H. (1979). Research related to the mathematical learning process. In International Commission on Mathematical Instruction (Ed.), *New Trends in Mathematics Teaching*, vol IV (pp199-213). Paris, France: UNESCO.

Bell, A.W. (1976). A study of pupils' proof-explanations in mathematics. *Educational Studies in Mathematics* 7, 23-40.

Bell, A.W. (1979). The learning of process aspects of mathematics. *Educational Studies in Mathematics* 10, 361-387.

Biggs, J.B. & Collis, K.F. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy*. New York, NY: Academic Press.

Burton, L. (1980). Problems and puzzles. *For the Learning of Mathematics* 1(2), 20-23.

Burton, L. (1984). *Thinking Things Through*. Oxford, UK: Blackwell.

Carroll, L. (1871). *Through the Looking Glass and What Alice Found There*. London, UK: Macmillan.

Case, R. (1985). *Intellectual Development*. New York, NY: Academic Press.

Christiansen, B. & Walther, G. (1986). Task and activity. In B. Christiansen, A.G. Howson, & M. Otte (Eds.), *Perspectives on Mathematics Education* (pp243-307). Dordrecht, Holland: Reidel.

Davis, R.B. (1984). *Learning Mathematics: The Cognitive Science Approach to Mathematics Education*. London, UK: Croom-Helm.

Depuis, C. (1985). Problem-situation. In A.Bell, B.Low & J. Kilpatrick (Eds.), *Theory, Research and Practice in Mathematical Education* (pp447-450). Nottingham, UK: University of Nottingham Shell Centre for Mathematical Education.

Engel, A. (1985). *Elementary Mathematics from an Algorithmic Standpoint* (F.R. Watson trans.). Keele, UK: Keele Mathematical Education Publications.

Galbraith, P. (1981). Aspects of proving: A clinical investigation of process. *Educational Studies in Mathematics* 12, 1-28.

GAIM (Graded Assessment in Mathematics). (1988). *Development Pack*. Basingstoke, UK: Macmillan.

GAIM (Graded Assessment in Mathematics). (in press). *Complete Pack*. Basingstoke, UK: Macmillan.

Greeno, J.G. (1973). The structure of memory and the process of solving problems. In R.L. Solso (Ed.), *Contemporary Issues in Cognitive Psychology*: The Loyola symposium. Washington, D.C.: Winston.

Her Majesty's Inspectors of Schools (HMI). (1985). *Mathematics from 5 to 16*. London, UK: HMSO.

Hiebert, J. & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert & P. Lefevre (Eds.), *Conceptual and Procedural Knowledge:* The Case of Mathematics. Hillsdale, NJ: Erlbaum.

Horton, B. (1979). *Tests of Mathematical Process*. Nottingham, UK: University of Nottingham Shell Centre for Mathematical Education.

Katona, G. (1942). Organising and memorising: A reply to Dr. Melton. *American Journal of Psychology* 55, 273-275.

Maier, N.R.F. (1945). Reasoning in humans III: The mechanisms of equivalent stimuli of reasoning. *Journal of Experimental Psychology* 35, 349-360.

Mason, J. (1984). *Learning and Doing Mathematics*. Milton Keynes, UK: Open University Press.

Mason, J., Burton, L. & Stacey, K. (1982). *Thinking Mathematically*. London: Addison-Wesley.

Mayer, R.E. (1983). *Thinking, Problem Solving and Cognition*. New York, NY: Freeman.

Mellin-Olsen, S. (1987). *The Politics of Mathematics Education*. Dordrecht, Holland: Reidel.

Novak, J.D. (1986). The importance of emerging constructivist epistemology for mathematics education. *Journal of Mathematical Behaviour* 5(2), 181-184.

OCEA (Oxford Certificate of Educational Achievement). (1987a). *Mathematics: General Reference*. Oxford, UK: Oxford International Assessment Services Limited.

OCEA (Oxford Certificate of Educational Achievement). (1987b). *Mathematics: Putting it into Practice*. Oxford, UK: Oxford International Assessment Services Limited.

Pascual-Leone, J. (1970). A mathematical model for the transition rule in Piaget's developmental stages. *Acta Psychologica* 32, 301-345.

Piaget J. (1956) Les stades du development mentale chez l'enfant et l'adolescent. In P. Osterreich, J. Piaget, R de Saussure, J.M. Tanner, H. Wallon and R. Zarro (Eds.), *Le Probleme des Stades en Psychologie de l'Enfant* (pp33-49). Paris, France: Presse Unversi taire de France.

Piaget J. (1978). *Success and Understanding*. Cambridge, MA: Harvard University Press.

Polya, G. (1945). *How to Solve It*. Princeton, NJ:Princeton University Press.

Scheffler, I. (1965). *Conditions of Knowledge: An Introduction to Epistemology and Education*. Chicago, IL: University of Chicago Press.

Schoenfeld, A.H. (1985). *Mathematical Problem Solving*. New York, NY: Academic Press.

Secondary Examinations Council (SEC) Mathematics Draft Grade Criteria Working Party (1985). *Report*. London, UK: Department of Education and Science.

Skemp, R.R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching* 77, 20-26.

SMILE (Secondary Mathematics Independent Learning) (1975). *Class Pack of Cards 0301-0500*. London, UK: Inner London Education Authority Learning Materials Service.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving and W. Donaldson (Eds.), *Organisation of Memory*. New York, NY: Academic Press.

Van Hiele, P.M. (1986). *Structure and Insight: A Theory of Mathematics Education*. New York, NY: Academic Press.

Wells, D.G. (1986). *Problem Solving and Investigations*. Westbury-on-Trym, UK: Rain Publications.

Wertheimer, M. (1959). *Productive Thinking*. New York, NY: Harper & Row

Preparation of the Document

Camera-ready copy of this publication was produced on an Apple LaserWriter II NT™ printer in the Faculty of Education at the University of British Columbia. The text of the papers is printed in 10-point Goudy with titles in Stone. Papers were received from the authors in a variety of formats and transferred, first into Microsoft Word™ for the Macintosh™ microcomputer, and then into Aldus Pagemaker™ for setting up the pages. Formulas and other mathematical expressions and symbols were produced in MathType™ which is published by Design Associates, Inc. Some of the diagrams and figures were pasted in.